Nature Communications

https://doi.org/10.1038/s41467-025-67280-9

Article in Press

Medicago super-pangenome reveals adaptive advantages and evolutionary constraints in autotetraploid alfalfa

Received: 5 February 2025

Accepted: 26 November 2025

Cite this article as: Zhang, F., Wei, C., Shi, X. *et al. Medicago* super-pangenome reveals adaptive advantages and evolutionary constraints in autotetraploid alfalfa. *Nat Commun* (2025). https://doi.org/10.1038/s41467-025-67280-9

Fan Zhang, Chunxue Wei, Xiaoya Shi, Shuo Cao, Xiaodong Xu, Zhiyao Ma, Yanling Peng, Rida Arshad, Hui Xue, Zhen Zhang, Wei Zhang, Yanshuai Xu, Yang Dong, Lianzhu Zhou, Xuejing Cao, Mengrui Du, Xu Wang, Zhiwu Zhang, Ruicai Long, Junmei Kang, Yongfeng Zhou & Qingchuan Yang

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

ARTICLE IN PRESS

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

ARTICLE IN PRESS

Medicago super-pangenome reveals adaptive advantages and evolutionary constraints in autotetraploid alfalfa

Fan Zhang[1,2§], Chunxue Wei[1§], Xiaoya Shi[2,3§], Shuo Cao[2], Xiaodong Xu[2], Zhiyao Ma[2], Yanling Peng[2], Rida Arshad[2], Hui Xue[2], Zhen Zhang[2], Wei Zhang[4], Yanshuai Xu[2], Yang Dong[2], Lianzhu Zhou[2], Xuejing Cao[2], Mengrui Du[2], Xu Wang[2], Zhiwu Zhang[5], Ruicai Long[1], Junmei Kang[1*], Yongfeng Zhou[2,4*], Qingchuan Yang[1*]

[1]Institute of Animal Science, Chinese Academy of Agricultural Sciences, Beijing, 100193, China
[2]State Key Laboratory of Tropical Crop Breeding, Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Key Laboratory of Synthetic Biology, Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, Guangdong, 518000, China
[3]College of Enology, Heyang Viti-Viniculture Station, Ningxia Helan Mountain's East Foothill Wine Experiment and Demonstration Station, Northwest A&F University, Yangling, Shaanxi, 712100, China
[4]State Key Laboratory of Tropical Crop Breeding, Tropical Crops Genetic Resources Institute, Chinese Academy of Tropical Agricultural Sciences, Haikou, Hainan, 571101, China
[5]Department of Crop and Soil Sciences, Washington State University, Pullman, WA, 99163, USA

[§]Fan Zhang, Chunxue Wei and Xiaoya Shi contributed equally to this work.

[*]Corresponding authors: Junmei Kang (kangjunmei@caas.cn), Yongfeng Zhou (zhouyongfeng@caas.cn), Qingchuan Yang (yangqingchuan@caas.cn).

ARTICLE IN PRESS

Abstract

ARTICLE IN PRESS

The genetic basis for the adaptive advantages of polyploids over their diploid relatives remains poorly understood. To address this knowledge gap, we generate a haplotype-resolved autotetraploid alfalfa (*Medicago sativa* subsp. *sativa*) genome and construct a super-pangenome from 13 genomes across seven *Medicago* taxa. We discover substantial gene content variation in alfalfa, with only 20.1% of genes present on all four haplotypes. Within this group, 53.3% are core genes conserved across the *Medicago* genus, which we term 'tetra-copy core genes'. We find these genes are significantly enriched in climate-adaptation-associated genes (1.60-fold) and stress-responsive differentially expressed genes (1.61-fold). Paradoxically, they also carry a high genetic burden, with 80.1% of deleterious variants located in coding regions. Indeed, overexpressing a representative tetra-copy core gene, the glycine decarboxylase (*MsGDC*), improves both biomass and nitrogen use efficiency, despite its high genetic burden. Our study reveals the trade-off between adaptation and evolutionary constraints mediated by tetra-copy core genes, facilitating polyploid genetics and alfalfa breeding.

ARTICLE IN PRESS

Introduction

Polyploidy, or whole genome duplication (WGD), has been considered an important evolutionary driving force in plant evolution, impacting from the past to the present and influencing levels from the molecular scale to the ecological level[1,2]. Polyploidy is estimated to have occurred in 30% to 70% of angiosperms, either as currently existing polyploids or as ancient polyploidies (paleo-polyploidies)[3-5]. This process encompasses gene duplication, gene loss, genetic exchange among sub-genomes, and involves molecular and physiological regulation, as well as gene expression and epigenetic regulation[4,6,7]. WGD provides genetic resources for evolutionary innovation and adaptation, enhancing fitness and phenotypic diversity. Compared to their diploid progenitors, polyploids tend to have larger organism size, enhanced competitive advantages, and higher fitness levels. Consequently, they often exhibit greater adaptability to environmental changes and occupy broader ecological niches[3,4,8-10].

Polyploids exhibit four main potential adaptive strategies. (1) In the short term, polyploids can mask deleterious mutations because recessive harmful variants are effectively hidden in a polyploid population, reducing genetic burden. This provides newly formed polyploids with immediate advantages, allowing them to establish niches. However, these advantages may diminish over time due to the accumulation of mutations[3,5]. (2) Adaptive processes may become more efficient after WGD due to an increase in effective population size and a higher rate of beneficial mutations, leading to enhanced selective efficiency and fitness[4,5,11,12]. (3) Duplicated genes offer genetic material for the evolution of new functions (neofunctionalization) or more specific functions (sub-functionalization). This process can lead to significant biological innovations, facilitating adaptation to new niches[13-16]. (4) Changes in gene expression and epigenetic patterns can alter physiology, metabolism, and morphology, thereby enhancing adaptive abilities[4,7,9,15]. While polyploidy is associated with an increase in adaptability, it is also regarded as an evolutionary 'dead-end'. Several theories attempt to explain this fate, including the accumulation of deleterious variants[4,5,17], lower diversification rates and higher extinction rates compared to closely related diploids[18,19], highly dynamic genomes[3,12,20] and abnormal meiosis with multivalent pairing among three or more homologous chromosomes[6,13]. Yet, the evidence for these hypotheses is largely derived from paleo-polyploids, forcing inferences to be drawn from the complex genomic signatures of ancient events. Furthermore, current research on polyploids predominantly prioritizes allopolyploids, which involve both WGD and hybridization effects. This

ARTICLE IN PRESS

combination obscures the individual contributions of polyploidy and hybridization, complicating our understanding of how polyploidy alone affects adaptability. Therefore, studying autopolyploids, which result from WGD within a single species, eliminates the confounding effects of hybridization, making them more suitable for investigating how evolutionary adaptability is influenced by WGD[3,13].

The construction of pangenome provides a powerful framework to analyze patterns of gene retention and loss, shedding light on preferences for gene conservation that relate to adaptability. For instance, core gene families were involved in fundamental biological functions in both cicer and arabidopsis[21,22]. Private genes play a crucial role in adaptation to local environment and climate conditions in populus[23]. Dispensable and private genes were enriched for abiotic and biotic-related genes in soybean[24]. Moreover, previous studies have demonstrated that gene retention following WGD is a nonrandom process[4,9]. The unique patterns of duplicated gene retention are correlated with fitness and adaptability[4,9,25]. The duplication of genes can create redundancy, which serves as a foundation for genetic and phenotypic innovation and can be utilized for adapting to environmental changes[9,25]. However, the identification of duplicated genes in polyploids has largely relied on gene family tree[1,26] or synonymous substitution rates (Ks values)[27,28], which are influenced by divergence time and nucleotide substitution rates. With advancements in sequencing and assembly strategies, the identification of allelic copy number variation, which results from gene duplication and retention following WGD, in autopolyploids can now be achieved with greater accuracy using phased haplotype-resolved genomes[29-31]. Therefore, a comprehensive approach that integrates super-pangenome analyses with phased autopolyploid genomes, uncovering gene retention over long evolutionary periods and allelic imbalance, provides a powerful framework for an in-depth understanding of adaptive mechanisms in polyploids[32,33].

The genus *Medicago*, a member of the legume family, is highly diverse. It contains 87 species with various ploidy levels (diploid, tetraploid, and hexaploid) that are adapted to a wide range of geographic distributions and climates. This genus is famous for its high protein production, nitrogen-fixing ability, and positive impact on soil ecology. Its members include diploid model plant *M. truncatula* and the autotetraploid high-quality forage alfalfa (*M. sativa* subsp. *sativa*,

ARTICLE IN PRESS

hereafter subsp. *sativa*). Compared to the diploid ancestor (*M. sativa* subsp. *caerulea*, hereafter subsp. *caerulea* or diploid alfalfa), the cultivated autotetraploid alfalfa presents higher yield, greater stress tolerance, improved climate adaptability, and a wider distribution range[34]. Understanding the genomic basis of these adaptive advantages has recently become feasible with the release of several *Medicago* genomes, including *M. truncatula*[35-37], *M. polymorpha*[38], *M. ruthenica*[39,40], subsp. *caerulea*[41] and alfalfa[42-44]. However, the absence of a high-quality phased tetraploid alfalfa genome makes it challenging to identify allelic copy number variation following WGD, thereby hindering our understanding of autopolyploid adaptability.

In this study, we generate a high-quality, haplotype-resolved autotetraploid alfalfa genome alongside a comprehensive super-pangenome for the *Medicago* genus using 13 genomes from seven taxa (six species, including two subspecies of *M. sativa*). Our analysis of gene retention reveals a distinct set of genes that we term 'tetra-copy core genes', which are defined as genes retained on all four allelic copies in alfalfa that are also core members of the *Medicago* genus. We demonstrate that these tetra-copy core genes are significantly enriched in climate-adaptation-associated genes and stress-related differentially expressed genes. Moreover, tetra-copy core genes exhibit higher overall expression than other genes, with one allelic copy showing significantly higher expression than the other three. Paradoxically, these functionally important genes also carry a disproportionately high genetic burden from deleterious mutations. We functionally validate this trade-off by showing that overexpression of a representative tetra-copy core gene, *MsGDC*, improves both biomass and nitrogen use efficiency, despite its high genetic load. Our study resolves a central paradox of polyploid adaptation by demonstrating how 'tetra-copy core genes' mediate a crucial balance between adaptative advantages and deleterious mutation accumulation. These findings provide a framework for understanding autopolyploid evolution and accelerating the genetic improvement of alfalfa.

Results

Genome assemblies and annotations of *Medicago* genomes

We sequenced two *M. sativa* subspecies (autotetraploid subsp. *sativa* and diploid subsp. *caerulea*) employing a combination of long reads sequencing platforms (PacBio and Nanopore sequencing) and high-throughput chromosome conformation capture sequencing (Hi-C) approaches. For

ARTICLE IN PRESS

autotetraploid alfalfa, we selected the high-yielding and salt-tolerant Chinese alfalfa cultivar Zhongmu-4 as the representative material. The Zhongmu-4 genome was sequenced, generating 109.84 Gb (~137×, compared to the ~800 Mb haploid genome of alfalfa) of high-fidelity (HiFi) long reads and 18.25 Gb (~23×) of Oxford Nanopore Technology (ONT) long reads. Combined with our previously sequenced 285 Gb (~356×) Hi-C data[44], we conducted the genome assembly (Supplementary Fig. 1). We initially generated contigs using HiFi and ONT reads with Hifiasm[45], and phased them into 32 chromosomes using Hi-C data, resulting in an assembly size of 2.54Gb (ZM4_V1.5, Supplementary Fig. 2). To enhance completeness and contiguity, we then divided the HiFi reads into 32 groups based on the haplotype information of our previously published ZM4_V1.0 and the ZM4_V1.5 genomes[44] (Supplementary Fig. 3). The genome was assembled into contigs using haplotype-resolved HiFi reads with Hifiasm, and subsequently improved to scaffold level using Ragtag (see Methods, Supplementary Fig.1).

Finally, we assembled ZM4_V2.0, a high-quality phased genome with a total size of 3.13 Gb across 32 chromosomes (contig N50 = 5.45 Mb). This assembly size is consistent with previous estimations for cultivated alfalfa and represents a significant improvement in quality over our previous version, ZM4_V1.0 (Supplementary Fig. 4)[44]. The raw HiFi and ONT data were mapped to ZM4_V2.0, achieving overall mapping rates of 99.93% and 99.92%, respectively. To establish a high-confidence set of alignments for assessing haplotype fidelity, we applied a MAPQ ≥ 10 filter, which retained 83.02% of HiFi reads and 88.34% of ONT reads. Critically, over 99% of the mapped HiFi reads aligned correctly to a single corresponding haplotype (Supplementary Fig. 5). Haplotype phasing analysis revealed a switch error rate of 1.97% and a Hamming error rate of 7.12%. This provides strong evidence for the high accuracy of our phasing framework, demonstrating that most haplotypic regions were successfully resolved. The Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis of the genome assembly, using the embryophyta_odb10 database (1614 orthologs), indicated a completeness score of 99.5%. Finally, using Merqury software[46], we assessed the base quality and genome completeness, resulting in scores of 61.88 and 97.8%, respectively (Supplementary Table 1). Therefore, the genome quality of ZM4_V2.0 is sufficiently high to serve as a reference genome for alfalfa and to support WGD analysis in autopolyploid. The sizes of four haplotype (hap) genomes were 791.79 Mb (contig N50 = 6.52 Mb) for hap1, 751.53 Mb (contig N50 = 4.98 Mb) for hap2, 809.40 Mb (contig N50 = 5.36

ARTICLE IN PRESS

Mb) for hap3 and 775.96 Mb (contig N50 = 5.45 Mb) for hap4. The four haplotypes showed significant gene collinearity (Supplementary Fig. 6) and strong genomic synteny with certain structural variations (SVs) (Supplementary Fig. 7). Additionally, a high collinearity was observed between the *M. truncatula* A17 (V5.0) and the ZM4_V2.0 genomes (Supplementary Fig. 8 and 9).

Using a homology-based, short and long-read transcriptomic annotation pipeline, we predicted 202,473 genes, and the number of genes was 51,000, 49,392, 50,958, and 51,123 for hap1, hap2, hap3, and hap4, respectively. The annotation was highly complete, with a BUSCO score of 99.8%, and 187,026 genes (92.4%) were functionally annotated according to the EggNOG database[47]. Taken together, ZM4_V2.0 represents a substantial improvement over the published alfalfa reference genomes[42-44].

Besides autotetraploid alfalfa *de novo* assembly, we also assembled a diploid subsp. *caerulea* genome employing 207.00 Gb PacBio CLR long reads, 52.75 Gb Illumina short reads, and 204.28 Gb Hi-C data. The assembled genome was 801.63 Mb (contig N50 = 1.61 Mb), with 96.4% BUSCO completeness, and the annotated gene number was 50,491. Furthermore, we annotated 47,963 genes for *M. lupulina* and 46,150 genes for *M. arabica* using the assembled genome provided by the Darwin Tree of Life Project (https://www.darwintreeoflife.org/)[48,49].

Genome evolution and gene conservation revealed by *Medicago* super-pangenome

Eleven published *Medicago* genomes were downloaded, and genome annotation was redone. Firstly, the total count, identity, and orientation of chromosomes were revised based on the model plant genome of *M. truncatula* A17 (5.0). Specifically, we revised the chromosome number of *M. polymorpha* from 7 to 8 (Supplementary Fig. 10). Secondly, gene annotation was performed using the same pipeline as that used for ZM4_V2.0. The number of annotated genes and the BUSCO assessment were improved compared to the raw version (Table 1). Using 13 *Medicago* genomes (19 haplotypes), we constructed a graph-based super-pangenome (Alfalfapan-V1.0) reference of seven *Medicago* taxa (Figure 1A). The Alfalfapan-V1.0 pangenome was constructed using the ZM4_V2.0 hap1 genome as a reference. It encompasses a size of 7.70 Gb across 8 chromosomes, with an average chromosome size of 968.82 Mb. The pangenome is 2.46-fold the size of the alfalfa ZM4_V2.0 genome, 9.72-fold that of the ZM4_V2.0 hap1 genome and 17.91-fold that of the *M.*

truncatula A17 genome (Figure 1D). We found that genome size in *Medicago* was highly correlated with transposable elements (TEs) content, indicating that TEs expansion is a primary driver of genome size evolution and divergence. Similarly, a comparable pattern of TEs contribution to genome expansion was observed in *Oryza*[50]. The two most abundant types of TEs, LTR Gypsy and LTR Copia, together contributed to ~50% of the total TEs content, underscoring their substantial role in the divergence of *Medicago* genomes (Figure 1E).

To investigate the evolution of gene content across the *Medicago* genus, we constructed a pangenome by identifying orthologous gene groups with OrthoFinder[51]. The gene based synteny analysis revealed high collinearity among *Medicago* species (Figure 1A). Furthermore, we examined a known translocation involving chromosomes 4 and 8 that distinguishes *M. truncatula* A17 from alfalfa[41,43,44]. Our multi-genome comparison reveals this translocation to be an accession-specific feature of the *M. truncatula* A17 genome, rather than a conserved, species-level difference. This highlights how a pangenome approach moves beyond simple pairwise comparisons to provide a more accurate and comprehensive view of SVs. Based on 5,106 single-copy orthologous genes of seven taxa, we constructed a phylogenetic tree of *Medicago* (Figure 1B). Our primary phylogenetic analysis resolved the overall species relationships among the studied *Medicago* taxa. In this tree, *M. truncatula* was identified as the sister group to the clade containing alfalfa and subsp. *caerulea*, while *M. ruthenica* was resolved as the earliest-diverging lineage. Interestingly, a more detailed analysis focusing specifically on the alfalfa haplotypes revealed a polyphyletic pattern (Supplementary Fig. 11). This topology suggests multiple independent origins of tetraploid alfalfa from genetically divergent diploid subsp. *caerulea* ancestors. Divergence time estimation placed the crown age of the studied *Medicago* clade at ~7.59 million years ago (Mya), with the tetraploid alfalfa lineage originating from its diploid progenitor ~2.13 Mya (Figure 1B).

A total of 69,767 orthologous gene families were identified from Alfalfapan-V1.0. Based on these gene families, the Alfalfapan-V1.0 super-pangenome is composed of 13,196 core (present in all genomes), 8,454 softcore (present in 11 and 12 genomes), 33,767 dispensable (present in two to ten genomes), and 14,350 private (only present in one genome) gene families (Figure 2A, Supplementary Table 2). The total number of pan-genes increased as more genomes were added,

ARTICLE IN PRESS

while the number of core genes steadily decreased (Figure 2B). Specifically, ZM4_V2.0 contained the highest number of orthologous gene families, totaling 48,151 (69.0% of pangenome). For the other genomes, the count of gene families ranged from 26,997 in Mpoly to 42,646 in XJDY (Figure 2C, Supplementary Table 3). It suggested that Alfalfapan-V1.0 contains at least 1.45 times more gene families (69,767 vs 48,151) than a single alfalfa genome. Interestingly, core gene patterns tend to cluster at the ends of chromosomes, while TEs are predominantly located in the pericentromeric regions. This suggests that core genes are more stable and resistant to TEs interference (Figure 2D).

The SVs were called for each genome using the ZM4_V2.0 hap1 as a reference. The results indicated that the *M. sativa-falcata* complex (including subsp. *sativa* and subsp. *caerulea*) exhibited a significantly higher number of SVs compared to other *Medicago* species (Figure 2E). This observation is consistent with the significantly lower number of mapped reads observed for the more distantly related species (Supplementary Fig. 12). Consequently, the true number of SVs in these related species was likely underestimated, highlighting the challenges of reliable SV calling at the genus level. Large SVs between species and haplotypes were often found in corresponding chromosomal regions. (Supplementary Fig. 13 and 14). Additionally, we generated population-level SVs using the PGGB and vg pipelines[52,53]. A total of 494,797 SVs were identified among the 13 genomes, with 208,045 SVs found exclusively in the *M. sativa-falcata* complex (Figure 2F). Using this comprehensive set of SVs, we further investigated the genetic relationships among these 13 genomes. The PCA results demonstrated clear separation among different species (Figure 2G). Interestingly, we detected divergence among the four alfalfa haplotypes across both the ZM4_V2.0 and XJDY (Figure 2H). However, the specific patterns of this divergence varied greatly at the individual chromosome level (Supplementary Fig. 15). Together, these findings demonstrate that SVs are powerful markers for resolving the complex and variable evolutionary histories of different sub-genomic regions in alfalfa.

The retention of core genes correlated with climate adaptation of alfalfa

Compared with the diploid ancestor subsp. *caerulea*, autotetraploid alfalfa has expanded into a wider range of climate regions and occupied broader ecological niches (Figure 1C). These results suggested that autopolyploidy plays a major role in the local adaptation of alfalfa. To confirm that

the allelic copy number variation originated from WGD without divergence, we carefully identified genes that displayed a 95.0% or higher protein sequence similarity across the four haplotypes in ZM4_V2.0 as allelic copies (Figure 3A, see Methods). The results indicated that a total of 109,715 (109,715/202,473, 54.2%) genes in ZM4_V2.0 could be considered as non-redundant genes (representing each gene only once despite allelic multiplicity). Among these, 49,344 (49,344/109,715, 45.0%) genes had one allelic copy, 23,095 (21.0%) had two allelic copies, 15,264 (13.9%) had three allelic copies, and 22,012 (20.1%) had four allelic copies (Supplementary Table 4). One representative genomic region on Chr1 was selected to illustrate allelic copy number variation (Figure 3B). Comparing the relationship between gene conservation and allelic copy number variance, we found that genes with four allelic copies constituted a substantial portion of the core genes (proportion: 53.3%; number:11,735; chi-squared test $p < 2.2e-16$), whereas genes with one allelic copy represented a substantial proportion of the dispensable genes (proportion: 41.7%; number: 20,562; chi-squared test $p < 2.2e-16$). (Figure 3C, Supplementary Table 5). It indicated that gene retention and loss following WGD were nonrandom. Analysis of selective pressure indicated that the log(Ka/Ks) value gradually decreased as the allelic copy number increased. Core genes with four allelic copies exhibited the lowest median Ka/Ks value (Figure 3D). The core genes that retained all four allelic copies in alfalfa represent a set of evolutionarily conserved duplicates within the *Medicago* genus; we therefore term them 'tetra-copy core genes'. In contrast, dispensable genes present as a one allelic copy in alfalfa exhibited a high median Ka/Ks ratio, suggesting they are undergoing rapid adaptive evolution, possibly as a mechanism for adaptation to new environmental pressures; we therefore assigned them the name 'unique-essential genes'. We developed a genome-wide gene classification framework by partitioning all 202,473 genes in ZM4_V2.0 into six groups based on gene conservation and allelic copy number variation. These groups include unique-essential genes (proportion: 10.2%; number: 20,562), one-copy genes (excluding unique-essential, 14.2%, 28,782), two-copy genes (18.4%, 37,344), three-copy genes (17.9%, 36,330), four-copy genes (excluding tetra-copy core genes, 17.7%, 35,873), and tetra-copy core genes (21.5%, 43,582) (Figure 3E, Supplementary Table 6). These six partitions were established as reference gene groups for various downstream analyses. For instance, they provide a genomic baseline proportion for assessing the enrichment of gene sets such as differentially expressed genes (DEGs) and candidates from genome scans for climate adaptation.

ARTICLE IN PRESS

ARTICLE IN PRESS

To examine the preferential retention and loss of genes related to adaptation following autopolyploidy, we further analyzed the contribution of each type of genes to local climate adaptation. At first, we examined the proportion of climate adaptation related genes within each of these six groups. The climate adaptation-associated genes were calculated by latent factor mixed model (LFMM) and population branch statistic (PBS) analysis[54]. A total of 2,429 genes were identified to be associated with climate adaptation. These genes were not evenly distributed among the six gene groups; instead, they were significantly enriched in the tetra-copy core genes. This single group accounted for 836 (34.4%) of the adaptation-associated genes, representing a 1.60-fold enrichment over the genomic baseline proportion of 21.5% (chi-squared test, $p < 2.2e-16$; Figures 3E and 4A). Our finding suggested that tetra-copy core genes play a significant role in climate adaptation.

To investigate the role of tetra-copy core genes in environmental adaptation, we analyzed RNA-seq datasets related to biotic/abiotic stress from 23 publications/research groups (covering 124 distinct expression conditions) (Supplementary Data 1 and 2). Our analysis focused on identifying both DEGs and genes ranked within the top percentiles of expression in each condition. Although the total number of DEGs varied significantly across treatments (mean = 689), the proportional contribution of six gene groups to the DEG sets remained remarkably consistent. The group of tetra-copy core genes was significantly overrepresented among the DEGs. The median proportion of these genes was 34.6%, a 1.61-fold enrichment over the genomic baseline proportion of 21.5% (one-sample Wilcoxon test, $p = 1.93e-4$; Figure 4B). Furthermore, unique-essential genes and one-copy genes had the highest fold change of DEGs (Supplementary Fig. 16). These results suggest that tetra-copy core genes contribute significantly to stress tolerance, while unique-essential genes are very susceptible to stress.

The level of gene expression is a key determinant of adaptive responses to environmental stress[55]. After analyzing six gene groups, we observed that genes with more allelic copies tended to have higher expression levels, and similar expression patterns were observed in autotetraploid potato[56]. Additionally, we found that the tetra-copy core genes had the highest median expression level (Figure 4C). Analysis of the four allelic copies within each non-redundant gene revealed a

ARTICLE IN PRESS

characteristic stepwise gradient, in which one allele consistently exhibited significantly higher expression than the other three (Figure 4D). This suggests that the high expression of a specific allele within the non-redundant gene plays an important role in maintaining the high dosage of gene products. In addition, we examined the proportion of each gene group among the top expressed genes (from top 50% to top 1%) under normal (pre-treatment), biotic, and abiotic stress conditions (Figure 4E). Across all conditions, we found that gene groups with four allelic copy numbers were disproportionately represented among top expressed genes. Specifically, the tetra-copy core genes were the most prominent, accounting for an average of 38.6% among top expressed genes across the various expression cutoffs (from top 50% to 1%). This represents a significant 1.80-fold enrichment over their 21.5% genomic baseline proportion. The four-copy genes were also consistently overrepresented (contributing an average of 21.1%), while the remaining four gene groups were underrepresented. Interestingly, the proportion of genes classified as unique-essential genes, one-copy genes, and two-copy genes increased among the top 1% of expressed genes, indicating their prominent role in the transcriptional regulation network. In summary, the tetra-copy core genes were consistently the most abundant and enriched group among top expressed genes, suggesting that these highly conserved genes play a key role in basic biological functions and stress adaptation.

To identify genes consistently expressed under various conditions, we counted the occurrences of highly expressed genes. We analyzed a total of 124 different conditions, including variations in accessions, treatments, and tissue types. For each condition, we defined highly expressed genes (HEGs) as the top 10% of expressed genes, with the number of HEGs ranging from 2,106 to 6,225 per condition (Supplementary Data 2). From this dynamic set, we identified a core group of 203 genes that were consistently highly expressed across more than 90% of conditions (Supplementary Fig. 17, Supplementary Data 3). Gene Ontology (GO) analysis revealed that these 203 genes were enriched in responses to water deprivation, cellular response to hypoxia, and reproductive processes. Among the 203 genes, we identified 81 that belong to the tetra-copy core genes (Supplementary Data 3). While a general pattern of unbalanced number of expressed genes was observed among the four haplotypes in both leaf and root tissues (with Hap4 highest and Hap1 lowest), our targeted analysis of this specific 81-gene subset revealed no significant differences in their expression level among the haplotypes (Supplementary Fig. 18A and B, Supplementary Table

7). Most non-redundant genes were expressed in only one haplotype, and none were simultaneously expressed in all four haplotypes (Supplementary Fig. 18C). This suggests that although all four allelic copies are conserved during evolution, their homology-dependent gene silencing mechanisms or genetic regulation prevent their simultaneous expression under stressful conditions[7,33,57,58].

Genetic burden of tetra-copy core genes and the long-term evolutionary fate of alfalfa

Our results showed that tetra-copy core genes are important for climate adaptation. However, the genetic burden of these conserved genes may accumulate over evolutionary timescales[3]. We calculated genomic sequence constraints using GERP score[59]. The results showed that ~1.5% (45.95 Mb) of the alfalfa genome was predicted as evolutionarily constrained sites (GERP score > 3) (Figure 5A). We further calculated the number of deleterious variants for each of the six gene groups (Supplementary Data 4). There are 6.41Mb (13.9%) deleterious variants located within the gene region (promoter + transcribed region). We found that the average number of deleterious variants increased with the rise in allelic copy number. The mean variants number is 65.25 for unique-essential genes, 53.21 for one-copy gene, 63.50 for two-copy genes, 97.64 for three-copy genes, 142.12 for four-copy genes and 198.89 for tetra-copy core genes. A significant proportion of the variants (63.6%) within the genes are derived from coding sequence (CDS) regions. Moreover, this pattern was consistent across genes with varying allelic copy numbers (Figure 5B).

Interestingly, the pattern of deleterious variants in unique-essential genes differed from that in other gene groups. Unlike other gene groups where CDS regions harbor the most deleterious variants, in unique-essential genes, the highest accumulation of deleterious variants was found in the 3' untranslated region (3'UTR). Tetra-copy core genes exhibited the highest number of deleterious variants among the six gene groups, with 80.1% of these variants located in the CDS regions. This result indicates that for every 1 kb genomic region, an average of 164 deleterious variants are located within CDS, indicating a high deleterious burden in tetra-copy core genes. We observed a consistent spatial distribution of tetra-copy core genes across the four haplotypes, with a significant enrichment clustering near the chromosomal ends (Supplementary Fig. 19), suggesting that deleterious variants are difficult to eliminate through recombination and selection.

ARTICLE IN PRESS

To compare the genetic burden among haplotypes, we quantified the number of deleterious variants within each allelic copy of the non-redundant genes. Our analysis of the entire genome revealed an average of 65.24 deleterious variants per gene. Using this value as a threshold, we then classified genes exceeding this average as deleterious genes. Our analysis revealed that 20,203 (46.4%) of the tetra-copy core genes (14,776 non-redundant genes) were identified as deleterious genes. Among them, there are 5,082 in hap1, 5,045 in hap2, 4,999 in hap3, and 5,077 in hap4. Furthermore, we assessed the distribution of deleterious alleles within each of the 14,776 non-redundant genes. Our analysis revealed that 5,070 (34.3%) of these non-redundant genes carried deleterious alleles on all four of their haplotypes (16,414 genes, Supplementary Data 5). In contrast, 4.9%, 5.2%, and 9.5% of the non-redundant genes carried deleterious alleles on only one, two, or three haplotypes, respectively (Figure 5C). Non-redundant genes carrying four deleterious allelic copies (5,070) were enriched in stress response, reproductive development, and protein ubiquitination (Figure 5D). It indicates that many key functional genes, including those environmental response genes, have accumulated a high genetic burden across all haplotypes. Specifically, our population-level analysis revealed that autotetraploid alfalfa has accumulated ~1.5-fold more deleterious variants than its diploid ancestor, subsp. *caerulea* (Figure 5E). This pattern was most pronounced in the tetra-copy core genes; while this group already had the highest genetic burden in the diploid ancestor, the burden was even greater in the autotetraploid alfalfa subspecies (Figure 5F). This finding suggests that WGD masked deleterious variants and increased the genetic burden, especially in tetra-copy core genes[3].

Overexpressed tetra-copy core gene increased adaptability in alfalfa

We selected the glycine decarboxylase gene (*MsGDC*) to validate the function of tetra-copy core genes, which is one of the 5,070 non-redundant genes where each of its four allelic copies carries a higher-than-average burden of deleterious variants. The *MsGDC* gene participates in photorespiration and plays a role in the utilization of CO_2 and NH_3[60]. We aim to determine whether the ability to utilize CO_2 or NH_3 can be enhanced and lead to phenotypic changes in overexpressed alfalfa. The comparative genomic results revealed that *MsGDC* is present in seven *Medicago* taxa, with four allelic copies found in the ZM4_V2.0 alfalfa genome (Figure 6A). The transcriptome results indicated that only one allelic copy of *MsGDC* was highly expressed, consistently across different accessions or lines (Figure 6B). Unexpectedly, only the highly expressed *MsGDC* (gene

ARTICLE IN PRESS

Msa125466 in ZM4_V2.0) showed a substantial increase in response to elevated CO_2 concentration[61]. The other three allelic copies exhibited only slight changes in response to shifts in CO_2 or NH_3 levels. This specific case, combined with our broader expression analyses (Figure 4D), suggests one allelic copy maintains the major roles in fundamental biological functions and stress tolerance, while the other three allelic copies may have acquired roles in genetic regulation[3].

Since the tetra-copy core genes exhibited the highest expression levels among all gene groups (Figure 4C) and one specific allele of *MsGDC* showed biased high expression (Figure 6B), we generated transgenic alfalfa with overexpressed *MsGDC* (Figure 6D, Supplementary Table 8). The results showed that the transgenic GS (the acronym for the gene *GDC* and the promoters ST-LS1) lines exhibited a significant increase in biomass (Supplementary Fig. 20). This indicates that enhancing the expression of even a single tetra-copy core gene can significantly increase alfalfa biomass, suggesting that employing these genes to improve alfalfa yield has substantial potential. The differences in promoters (G lines: 35S or GS lines: ST-LS1) significantly affected plant biomass. The GS lines exhibited greater weight and height compared to the wide-type (WT) and G lines (Supplementary Fig. 20), and we also observed significant differences in gene expression levels between the promoter lines (Supplementary Fig. 21). This highlights the critical importance of promoter selection for realizing the biomass-boosting potential of this gene, as demonstrated by the superior performance of the ST-LS1 promoter[62].

To further assess stress tolerance, we examined the *MsGDC*-overexpressing *M. truncatula* lines. Under normal nutrient conditions, these lines exhibited a phenotypic pattern similar to that of transgenic alfalfa, with the GS lines showing higher biomass than the WT (Figure 6E). In contrast, when subjected to nitrogen-free treatment, the GS lines maintained good survival rates, whereas the WT plants showed poor growth and survival (Figure 6F). Moreover, there was a significant difference in chlorophyll content between the GS lines and WT plants (Supplementary Fig. 22). These results suggest that *MsGDC* can enhance tolerance to nitrogen stress and improve nitrogen use efficiency in *M. truncatula*. Finally, we calculated the genetic burden of *MsGDC*. We found that most deleterious variants were present in the CDS regions at a remarkably high density. This observation was consistent for each of the four allelic copies, demonstrating that *MsGDC* carries a high genetic burden (Figure 6C). To summarize, *MsGDC* has the potential to enhance alfalfa's

ARTICLE IN PRESS

biomass and improve the nitrogen stress tolerance of *M. truncatula*. It could be utilized to boost the yield and develop climate-resilient alfalfa varieties. However, the high genetic burden should be carefully managed in alfalfa breeding programs.

Discussion

Overall, we *de novo* assembled high-quality genomes for both diploid subsp. *caerulea* and autotetraploid alfalfa, and constructed a super-pangenome of *Medicago*. Based on the super-pangenome and haplotype-resolved alfalfa genome, we discovered that the retention of core genes following WGD was nonrandom. Our results highlight that tetra-copy core genes are significant contributors to the adaptive potential of autopolyploids. However, the high deleterious burden observed in these tetra-copy core genes can jeopardize long-term adaptive advantage. Our findings offer substantial evidence linking evolutionary 'dead ends' with short-term adaptive potential, thereby enriching our understanding of the evolutionary processes in autotetraploids (Figure 7). Additionally, using alfalfa as a model organism, our findings provide a roadmap for improving autopolyploid crops by focusing on both the dosage of core genes and the expression of their most effective alleles.

The genomic basis underlying the adaptive potential of polyploidy has fascinated scientists over the past century. Previous studies have demonstrated that WGD can create gene redundancy, which can be harnessed to evolve new functions or subfunctions, thereby contributing to evolutionary innovation and environmental adaptation[9,14]. Diploid plant species like poplar and soybean often rely on their flexible dispensable and private genes for local adaptation[23,24]. Meanwhile, duplicated genes may be retained for providing dosage benefits and enhancing regulatory networks. This hypothesis is supported by evidence from the paleopolyploid *Sonneratia*, where the retention of duplicated genes was selected in response to new environments[27]. This mechanism is particularly potent because even highly conserved core genes, typically viewed as functionally constrained, can serve as a direct source of adaptive material. Their duplication can immediately bolster the output of fundamental pathways such as signaling, transport, and metabolism, providing a rapid route to enhanced fitness[25]. Our results proved that tetra-copy core genes, rather than unique-essential genes, play a major role in climate adaptation and stress tolerance. The tetra-copy core genes, along with the retention, loss and enrichment of other climate-related genes, enhanced the

physiology, metabolism, and morphological resilience of alfalfa, thereby improving its adaptability[9]. Prospectively, if the phased pangenome of the *M. sativa-falcata* complex is released in the near future, we can further investigate how specific breeding processes (such as crossing and inbreeding) induce variants in tetra-copy core genes, ultimately yielding valuable insights into line-specific adaptability[63].

The four allelic copies of non-redundant genes typically exhibited a quantitatively graded expression pattern, in which one allele consistently showed significantly higher expression than the other three (Figure 4D). Similarly, in potatoes, duplicated genes also showed preferential allele expression[33]. This suggested that the most significant impact of WGD might be the emergence of new regulatory patterns for genes, rather than functional diversification or dosage effects. Specifically, WGD might lead to a rebalancing of duplicated gene dosage during transcription[3,14]. Previous research showed that functionally redundant paralogs exhibit higher genetic interactions than diverged paralogs[57]. The structural or functional entanglement of redundant paralogs may restrict gene expression. At the same time, the duplication of allelic genes could create novel allelic combinations and alter gene expression patterns. The unexpressed genes could achieve balance by regulating the dosage of expressed genes[13,14]. Alternatively, the retention of all duplicated copies may be important for creating an integrated regulatory network. Within this network, one specific allele is maintained at a high expression level to perform the primary biological function, while the other three allelic copies act primarily in a regulatory capacity to ensure its stable and robust expression. Therefore, we propose that the strategy of preferential allelic expression of tetra-copy core genes represents one specific regulatory pattern in autopolyploids[12].

WGD leads to massive gene duplication and genome-wide genetic redundancy. While this redundancy can provide evolutionary novelty, it also allows most duplicated genes to accumulate deleterious variants. This challenge is particularly acute in autopolyploids, where purifying selection is less efficient, leading to the progressive accumulation of deleterious variants[17,64]. Our findings in alfalfa are consistent with this phenomenon. We found that within its relatively recent WGD (~2 Mya), 5,070 non-redundant genes (16,414/202,473, 8.1% of the overall gene set) have accumulated a high load of deleterious variants, which are difficult to purge via selection (Supplementary Figure 19). The functional significance of this genetic load is profound; the

continued accumulation of deleterious variants can lead to gene function loss and abnormalities in meiotic segregation, ultimately threatening survival[3,6]. Indeed, we attempted to knock out the *MsGDC* gene, but this resulted in seedling lethality prior to transfer from the culture medium, confirming its essentiality. Taken together, this widespread accumulation of deleterious variants in thousands of tetra-copy core genes likely constrains the long-term adaptive advantages of alfalfa, lending strong support to the long-standing 'evolutionary dead-end' hypothesis for polyploids[4,19].

The accumulation of deleterious variants can also promote short-term adaptation in polyploids by providing a rich pool of genetic variation for selection to act upon[65]. For example, two convergent *de novo* mutations affect the function of the *TPC1* gene and are correlated with adaptation to serpentine soils[66]. A transposable element insertion disrupted the expression of the *FLC* gene, resulting in adaptive early flowering[67]. Furthermore, the insertion of new genetic variants can enhance the adaptability of autopolyploid species, as exemplified by the positive selection of transposable element insertions in key genes during the local adaptation of autotetraploid *A. thaliana*[67]. Additionally, ploidy-specific structural variations have increased the adaptive potential in autopolyploid *Cochlearia*[64]. Our findings in alfalfa clearly illustrate this evolutionary paradox. We found that while a significant fraction (34.3%) of tetra-copy core genes carry a high deleterious burden, they remain functionally important for alfalfa's adaptation. This highlights how polyploids can tolerate a high genetic load in the short term, potentially co-opting the underlying variation for immediate adaptive benefit.

The deleterious variants among tetra-copy core genes could negatively impact plant performance during the breeding process, as recombination might reveal recessive deleterious variants that threaten gene function or plant survival. This hypothesis was validated by the observation of significant inbreeding depression in alfalfa during self-pollination[68,69]. A creative strategy involves initially breeding diploid alfalfa with minimal deleterious variants by design, followed by inducing genome doubling to create tetraploid alfalfa. A similar breeding strategy has been explored in self-incompatible diploid potato. By carefully managing deleterious variants, it is possible to produce inbred lines and design ideal potato haplotypes[70,71]. In addition, beneficial variants for agronomic traits in crops like grapevine, potato, and tomato can be identified and utilized through GWAS and

ARTICLE IN PRESS

genomic selection[72-75]. Finally, tetraploid alfalfa can be developed through artificial polyploidy, similar to how tetraploid tomato was achieved using unreduced diploid gametes[76]. These studies provide effective methods to manage deleterious variants and design ideal alfalfa cultivars.

Methods

Genome sequencing and assembly

Genomic DNA was extracted from young leaves of the previously described Zhongmu-4 alfalfa plant[44]. DNA libraries were constructed using standard protocols for PacBio and Nanopore sequencing. Four PacBio Sequel II HiFi libraries were prepared to generate consensus reads. Additionally, a 100 kb Oxford Nanopore Technologies (ONT) library was created to obtain ultra-long reads. The Hi-C data were obtained from our previous sequencing of Zhongmu-4[44]. Full-length transcriptomic sequences were generated using mixed samples from leaves, buds, and flowers.

The haplotype-resolved assembly of Zhongmu-4 was generated in three steps. First, HiFi and ONT reads were used to produce primary contigs using Hifiasm (v0.18.5-r499)[45]. Hi-C reads were aligned to the primary contigs using the Juicer (v2.0) pipeline[77]. The resulting contact maps were then used with the 3D-DNA (v.201008) pipeline for an initial round of automated scaffolding[78]. This draft assembly was loaded into Juicebox (v1.11.08) for manual inspection and correction of large-scale scaffolding errors. To finalize the assembly, the manually curated scaffold list was then used as input to re-run the 3D-DNA pipeline, which generated the haplotype-resolved pseudo-chromosomes (ZM4_V1.5). Second, the genomic continuity and completeness were further improved based on previous version of Zhongmu-4 genome (ZM4_V1.0)[44]. We mapped all HiFi reads to the ZM4_V1.5 genome, and kept reads with a mapping quality score (MAPQ) of more than 55. Reads that had a mapping quality score below 55 were subsequently remapped to the ZM4_V1.0 genome. Based on the collinearity between ZM4_V1.0 and ZM4_V1.5, the remapped reads were assigned to the corresponding chromosome of ZM4_V1.5 genome. A total of 89.5% of HiFi reads were assigned to 32 pseudo-chromosomes. Third, Hifiasm (v0.19.7-r598) was used to regenerate haplotype-resolved contigs by utilizing HiFi and ONT reads, referencing the strategy of genome assembly of potato C88[29]. The parameters used were '-D 10 -5 mapped_high_quality_MQ55_hifi_reads.txt --n-hap 4 --hom-cov 120'

(https://github.com/baozg/Potato_C88). As a result, we generated four groups of phased haplotype-resolved contigs (Supplementary Fig. 1). For each group, the contig reads were oriented and ordered to the chromosome-level genome based on the hap1 of ZM4_V1.5 using RagTag (v2.1.0) with default parameters[79,80]. The resulting chromosome-level genome was named ZM4_V2.0 and used for subsequent analysis. The genome completeness and base quality were evaluated using Merqury (v1.3)[46]. Phasing accuracy was evaluated by calculating switch and Hamming error rates using hap1 as the reference with a custom Python script, which compared a FreeBayes (v1.3.8)[81] derived variant set against a ground truth variant set established by MUMmer4 (v4.0.1)[82].

For subsp. *caerulea*, young leaves from one accession (PI 464714) were used for sequencing. The sequencing libraries for PacBio CLR and Hi-C were constructed according to the manufacturer's protocols. In total, we generated 207.00 Gb PacBio CLR long reads, 52.75 Gb Illumina short reads, and 204.28 Gb Hi-C data. The PacBio CLR reads were first assembled into contigs using Canu (v1.8)[83] and polished with Illumina short reads using Pilon (v1.23)[84]. The contigs were then aligned and anchored using Juicer (v1.5)[77] based on Hi-C data and subsequently assembled into chromosomes using 3D-DNA (v.180419)[78].

Annotation of genes and transposable elements

The gene annotation was performed using the GWAP pipeline (https://github.com/unavailable-2374/Genome-Wide-Annotation-Pipeline)[85]. In summary, the ZM4_V2.0 putative genes were identified using a combination of multiple data sources, including protein annotation information from the ZM4_V1.0 genome, full-length transcriptomic sequences, Illumina short-read RNA data from leaves, stems, and flowers, as well as protein sequences from *M. truncatula* (Mt_A17(V5.0))[35]. A preliminary gene model was constructed for the putative genes, followed by additional gene searches utilizing AUGUSTUS (v3.4.0)[86]. The identified putative gene fragments were then filtered through the following steps: removal of genes with duplicated regions, removal of genes with invalid CDS regions (where the end position is less than the start position), removal of genes with protein sequences containing a point sign or lengths shorter than 55 amino acids, and removal of genes not supported by any evidence. InterProScan (v5.56–89.0)[87] was utilized for the functional annotation of the retained genes. Furthermore, inconsistent annotation pipelines can

introduce significant biases, leading to inflated estimates of dispensable gene families[88]. To address this issue, we re-annotated the remaining 12 genomes using a consistent pipeline.

Transposable elements analysis was conducted based on the pangenome of *Medicago* using our established pipeline[85]. In brief, the *Medicago* repeat family database was initially constructed using RepeatModeler2 (open-2.0.3)[89] and subsequently identified with RepeatMasker (open-4.1.2)[90]. TEs families were identified using RepeatModeler2. Single-copy and erroneous annotations of TEs families were removed, and redundant sequences were further eliminated using NCBI-BLAST+ (v2.9.0). Unclassified repeat elements were identified with DeepTE[91]. Finally, the repetitive sequences were annotated using RepeatMasker.

Super-pangenome construction and gene family analysis

The *Medicago* graph pangenome was constructed using PGGB (v0.5.4)[52] with the ZM4_V2.0 hap1 genome serving as the reference. All 13 genomes, encompassing 19 haplotypes, were included in the construction of the graph pangenome. Single chromosomes of 19 haplotypes were merged in one input fa file. The parameter of PGGB was set as follows: "-s 5000 -l 25000 -p 90 -c 1 -K 19 -F 0.001 -g 30 -k 23 -f 0 -B 10M -n 19 -j 0 -e 0 -G 700,900,1100 -P 1,19,39,3,81,1 -O 0.001 -d 100". The VCF file was generated based on the gfa file of PGGB using vg with default parameter (v1.58.0)[53]. Structural variations were identified based on the VCF file results, which only include reference or alternate sequences exceeding 50bp in length. In order to achieve uniform comparisons among the 13 genomes, PCA analysis was conducted using SVs with no more than two alternate variants, and all alternate variants beyond the second were set to NA. The accession-specific SV was identified using SyRI (v1.5.4)[92], SVIM-ASM (v1.0.3)[93], and CuteSV (v2.1.2)[94], with the hap1 of ZM4_V2.0 as the reference, and was subsequently merged with Sniffles (v2.0.7)[95].

OrthoFinder (v2.5.4)[51] was utilized to identify orthologous gene families across 13 *Medicago* genomes using default parameters. GENESPACE (v1.2.3)[96] was employed to perform synteny analysis among these genomes. Collinear figures were generated using MCscan[97] and R (v4.3.2). Phylogenetic trees were constructed based on single-copy orthologs identified by OrthoFinder from seven genomes: *M. truncatula* (A17_5.0), *M. ruthenica* (Mru_zhiwusuo), *M. polymorpha*,

M. lupulina, *M. arabica*, subsp. *caerulea* (Mca_long), and ZM4_V2.0 (hap1). The tree inference was performed with raxml-ng (v1.1)[98]. The divergence time between different species was estimated using PAML (v4.9)[99] with default parameters. The calibration points were set at two specific points: the divergence time between the *M. sativa-falcata* complex and *M. truncatula* was 6~9 Mya, while the divergence time between the genus *Medicago* and *Trifolium pratense* was estimated to be between 13~24 Mya. The final phylogenetic tree was visualized and annotated in FigTree (v1.4.4) (https://tree.bio.ed.ac.uk/software/figtree/).

Gene conservation and expression analysis

To identify allelic copy number variation within haplotype-resolved ZM4_V2.0 genome, we performed an all-to-all BLASTP (v2.2.31)[100] analysis using the protein sequences from all four haplotypes. The allelic copy number of a given gene was then defined as the total number of its homologous alleles (including itself) found across the four haplotypes. The filtering parameters were set as follows: "e-value < 1e-10, pident > 95%, (1 - mismatch/length) > 95%, and Chr_query = Chr_match". For each gene, the top-matching sequence with the highest percentage identity was identified as its corresponding allelic copy. These identified pairs were then treated as a single non-redundant gene for all subsequent analyses. To focus on allelic variation across haplotypes, we treated tandemly duplicated genes on the same haplotype as individual entities. This approach allows members of a tandem array to sort into different non-redundant gene groups, with the result that the allelic copy number for any given group is a maximum of four. The allelic copy number of a given gene was defined as one plus the number of other haplotypes on which a homologous allele was identified (via all-to-all BLASTP analysis), resulting in allelic copy numbers ranging from one (no homologs on other haplotypes) to four (homologs on all three other haplotypes). Additionally, we distinguished paralogs from allelic copies by strictly defining the latter as the gene with the highest protein similarity across haplotypes. Furthermore, this protein-based definition differs from identifying genes via presence/absence variation (PAV), which is often linked to structural variation (SV). The key distinction is that PAV is more suited to identifying newly arisen/loss genes, whereas our method offers insights into the detailed evolution of established non-redundant genes.

ARTICLE IN PRESS

ARTICLE IN PRESS

We determined the conservation status of each orthologous gene family (orthogroup) identified by OrthoFinder. Using the Statistics_Overall.tsv file, families were categorized as core (present in all genomes), softcore (present in 11-12), dispensable (present in 2-10), or private (present in only one). Subsequently, this family-level conservation status was assigned to every individual gene within its respective orthogroup by parsing the Orthogroups.tsv file. The relationship between gene conservation status (i.e., core vs. dispensable) and allelic copy number was analyzed using the R package dplyr. Based on this analysis, core genes with four allelic copies were defined as 'tetra-copy core genes', while one allelic copy dispensable genes were classified as 'unique-essential genes'. The remaining genes were categorized based on their respective allelic copy numbers. For detailed steps, please refer to the code on GitHub (https://github.com/futurefanzhang/genomic-adaptation-of-autotetraploid-alfalfa/tree/main/step3).

The climate adaptation-associated genes were analyzed based on the VCF file generated in our previous study[54] (https://ngdc.cncb.ac.cn/gvm/getProjectDetail?Project=GVM000965). The reference transfer was conducted using the following steps: First, the genome sequences of haploid ZM4 were mapped to the ZM4_V2.0 hap1 genome using minimap2 (v2.21-r1071)[101]. Next, chain files were created using the minimap2chain function of transanno software (v0.4.5) (https://github.com/informationsea/transanno). Finally, the VCF file was generated with ZM4_V2.0 hap1 as the reference through the liftvcf function of transanno. The analyzed sample consisted of 240 accessions exhibiting a clear geographic clustering. The LFMM and PBS analyses were conducted using the same pipeline as that reported in our previous study[54]. In brief, we performed a genotype-environment association analysis using the univariate LFMM in the R package LEA (v.3.11.4)[102]. The environmental data for 19 bioclimatic variables (resolution 2.5 arcmin) were downloaded from the WorldClim website[103]. Local adaptation regions and candidate genes were detected by PBS analysis using PBScan[104], which involved grouping samples based on their geographic origin. Additionally, population-related analyses between subsp. *caerulea* and subsp. *sativa* were conducted using this VCF file.

The potential distribution regions of alfalfa and subsp. *caerulea* were inferred using a random forest model[54]. Briefly, this model was trained with the geographic coordinate data of accessions obtained from the USDA-GRIN database. To model the potential species distribution, we used a

set of bioclimatic variables (BIO2, BIO8, BIO9, BIO15, BIO17, BIO18, and BIO19) obtained from the WorldClim version 2.1 database at a spatial resolution of 2.5 arc-minutes[103]. The final distribution map was generated by plotting the prediction raster from the model in R (v4.3.2) using the 'raster' and 'dismo' packages[105,106].

A total of 23 Bioproject RNA-seq datasets related to biotic/abiotic stress were used for gene expression analysis in this study (22 Bioproject downloaded from NCBI, one generated from this study). The materials, repeat information, treatments, experimental design, and publication details are provided in Supplementary Data 1. For each of 23 datasets, the RNA-seq raw data were initially mapped to the ZM4_V2.0 genome using Tophat2 (v2.1.1)[107]. The resulting BAM files were sorted using Samtools (v1.9)[108], and gene counts were calculated with HTSeq-count (v0.9.1)[109]. Gene expression FPKM values were calculated using a custom Perl script. Differentially expressed genes (DEGs) were identified using the DESeq2 R package (v1.42.0)[110], with thresholds of $|\log2(\text{FoldChange})| > 1$ and an adjusted *P*-value (FDR) < 0.05. A gene was considered expressed if its FPKM value was greater than zero. For each study, DEGs were identified through pairwise comparisons between samples. Subsequently, only the DEG sets from comparisons yielding more than 100 genes were retained. These selected sets were then used for gene groups enrichment analysis to mitigate potential statistical bias associated with small sample sizes. The corresponding information for DEGs and gene groups was merged using the dplyr package in R (v4.3.2). Gene Ontology analysis was performed using the DAVID website (https://david.ncifcrf.gov/)[111] and visualized using the ggplot2 R package[112]. Statistical analyses were performed to identify significant enrichments of gene groups. To test for associations between gene groups (e.g., tetra-copy genes) and conservation categories (e.g., core genes), we used a chi-squared test of independence based on 2x2 contingency tables. One row of the table contained the counts for the gene list of interest, while the second row contained the corresponding counts from the genomic baseline. For analyzing the proportion of specific gene groups within the DEGs across the 23 datasets, a one-sample Wilcoxon signed-rank test was employed. This non-parametric test was used to assess whether the median of the observed proportions was significantly different from a specific genomic baseline proportion (https://github.com/futurefanzhang/genomic-adaptation-of-autotetraploid-alfalfa/tree/main/step5).

ARTICLE IN PRESS

ARTICLE IN PRESS

Genetic burden analysis

The constrained sites in the ZM4_V2.0 genome were identified using the GERP++ software[59], with the rejected substitutions value considered as the GERP score. Multiple whole-genome sequence alignments were performed using LASTZ (v1.04.15)[113]. A total of 15 angiosperm genomes were included in the alignment, specifically: grape[85], poplar[114], *Arabidopsis thaliana*[115], red clover[116], soybean[117], pea[118], chickpea[119], *Arachis duranensis*[120], *M. truncatula* (A17_5.0)[35], *M. ruthenica*[40], *M. polymorpha*[38], *M. lupulina*[49], *M. arabica*[48], subsp. *caerulea* (Mca_long), and alfalfa (ZM4_V2.0). The alignment files were merged using multiz (v11.2) and converted to FASTA format using maf2fasta (v3)[121]. A phylogenetic tree was constructed with single copy genes using raxml-ng (v1.1)[98], and branch lengths modified based on fourfold degenerate sites using msa_view (v1.4) and phyloFit (v1.4) (http://compgen.cshl.edu/phast/phyloFit-tutorial.php). Finally, GERP scores were calculated for each position using the gerpcol function of GERP++. From this output, we first excluded all uninformative sites (those with a GERP score of zero). We then defined deleterious variants as those occurring at sites with a GERP score > 3 (rejected substitutions > 3). The detailed steps can be found on GitHub (https://github.com/futurefanzhang/genomic-adaptation-of-autotetraploid-alfalfa/blob/main/step4/gerp_score.sh).

Validation of *MsGDC* in transgenic experiment

The plant materials for treatment were derived from a single plant of the alfalfa cultivar Zhongmu No.1 (Zhongmu1) and a single plant of *M. truncatula* (R108). The full-length CDS sequence of the candidate gene *MsGDC* was cloned from the Zhongmu1 genome[43]. The stem-leaf-specific expression promoter ST-LS1 sequence was amplified and cloned based on the sequence from NCBI GenBank (accession No. X04753.1)[122]. The correctly sequenced vectors expressing *MsGDC*, driven by the 35S and ST-LS1 promoters, were transformed into *Agrobacterium tumefaciens* competent cells. Alfalfa leaves were transformed via *Agrobacterium tumefaciens* infection, and the transgenic seedlings were screened on a solid medium containing glyphosate (4 $mg \cdot L^{-1}$) through tissue culture. The expression level of *MsGDC* among its four allelic copies was calculated using a publicly available RNA-seq dataset[61].

ARTICLE IN PRESS

Two-month-old Zhongmu1 plants overexpressing *MsGDC* were grown in an artificial climate chamber (16 hours light at 24˚C, 8 hours dark at 22˚C, 60% humidity). We quantified leaf *MsGDC* expression via real-time PCR (7300 Real Time PCR System, ABI) and measured several growth-related phenotypes, including plant height, fresh weight, and dry weight. Three biological replicates were conducted for each test, with one plant selected for each replicate.

One-month-old wild-type (WT) and transgenic *M. truncatula* seedlings overexpressing *MsGDC* were selected and subjected to a nitrogen-free nutrient solution for an additional month. Plant phenotypes were examined, and 0.1 g of leaves were collected for chlorophyll extraction. Leaves from the N-free *M. truncatula* plants were immersed in 10 mL of 95% ethanol and kept in the dark for 48 hours until they were completely discolored. Absorbance values at 665 nm and 649 nm were measured using a NanoPhotometer instrument (Germany). Chlorophyll a, chlorophyll b, and total chlorophyll contents were subsequently calculated[123].

Data availability

All raw sequence data were uploaded to the National Genomics Data Center (NGDC) under BioProject PRJCA031790 [https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA031790] and NCBI Sequence Read Archive Bioproject PRJNA1179654 [https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1179654]. The genome assembly and annotation data have been deposited in the Figshare repository [https://doi.org/10.6084/m9.figshare.30556160.v1]. Source data are provided with this paper.

Code availability

The codes used in this study are available on GitHub [https://github.com/futurefanzhang/genomic-adaptation-of-autotetraploid-alfalfa]. Inquiries about the code and software usage can be directed to Fan Zhang (zhangfan06@caas.cn or zfan887@gmail.com).

References

1. Jiao, Y. *et al.* Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**, 97-100 (2011).

ARTICLE IN PRESS

2. Adams, K.L. & Wendel, J.F. Polyploidy and genome evolution in plants. *Current Opinion in Plant Biology* **8**, 135-141 (2005).
3. Baduel, P., Bray, S., Vallejo-Marin, M., Kolář, F. & Yant, L. The "Polyploid Hop": shifting challenges and opportunities over the evolutionary lifespan of genome duplications. *Frontiers in Ecology and Evolution* **6**, 117 (2018).
4. Van de Peer, Y., Mizrachi, E. & Marchal, K. The evolutionary significance of polyploidy. *Nature Reviews Genetics* **18**, 411-424 (2017).
5. Otto, S.P. The evolutionary consequences of polyploidy. *Cell* **131**, 452-462 (2007).
6. Comai, L. The advantages and disadvantages of being polyploid. *Nature Reviews Genetics* **6**, 836-846 (2005).
7. Cheng, F. *et al.* Gene retention, fractionation and subgenome differences in polyploid plants. *Nature Plants* **4**, 258-268 (2018).
8. Sattler, M.C., Carvalho, C.R. & Clarindo, W.R. The polyploidy and its key role in plant breeding. *Planta* **243**, 281-296 (2016).
9. Van de Peer, Y., Ashman, T.-L., Soltis, P.S. & Soltis, D.E. Polyploidy: an evolutionary and ecological force in stressful times. *Plant Cell* **33**, 11-26 (2021).
10. Ramsey, J. Polyploidy and ecological adaptation in wild yarrow. *Proceedings of the National Academy of Sciences* **108**, 7096-7101 (2011).
11. Otto, S.P. & Whitton, J. Polyploid incidence and evolution. *Annual Review of Genetics* **34**, 401-437 (2000).
12. Parisod, C., Holderegger, R. & Brochmann, C. Evolutionary consequences of autopolyploidy. *New Phytologist* **186**, 5-17 (2010).
13. Spoelhof, J.P., Soltis, P.S. & Soltis, D.E. Pure polyploidy: closing the gaps in autopolyploid research. *Journal of Systematics and Evolution* **55**, 340-352 (2017).
14. Panchy, N., Lehti-Shiu, M. & Shiu, S.-H. Evolution of gene duplication in plants. *Plant Physiology* **171**, 2294-2316 (2016).
15. Alix, K., Gérard, P.R., Schwarzacher, T. & Heslop-Harrison, J. Polyploidy and interspecific hybridization: partners for adaptation, speciation and evolution in plants. *Annals of Botany* **120**, 183-194 (2017).
16. Van de Peer, Y., Maere, S. & Meyer, A. The evolutionary significance of ancient genome duplications. *Nature Reviews Genetics* **10**, 725-732 (2009).

ARTICLE IN PRESS

ARTICLE IN PRESS

17. Monnahan, P. *et al.* Pervasive population genomic consequences of genome duplication in *Arabidopsis arenosa*. *Nature Ecology & Evolution* **3**, 457-468 (2019).
18. Arrigo, N. & Barker, M.S. Rarely successful polyploids and their legacy in plant genomes. *Current Opinion in Plant Biology* **15**, 140-146 (2012).
19. Mayrose, I. *et al.* Recently formed polyploid plants diversify at lower rates. *Science* **333**, 1257-1257 (2011).
20. Xiong, Z., Gaeta, R.T. & Pires, J.C. Homoeologous shuffling and chromosome compensation maintain genome balance in resynthesized allopolyploid *Brassica napus*. *Proceedings of the National Academy of Sciences* **108**, 7908-7913 (2011).
21. Khan, A.W. *et al.* Cicer super-pangenome provides insights into species evolution and agronomic trait loci for crop improvement in chickpea. *Nature Genetics* **56**, 1225–1234 (2024).
22. Lian, Q. *et al.* A pan-genome of 69 *Arabidopsis thaliana* accessions reveals a conserved genome structure throughout the global species range. *Nature Genetics* **56**, 982–991 (2024).
23. Shi, T. *et al.* The super-pangenome of *Populus unveils* genomic facets for its adaptation and diversification in widespread forest trees. *Molecular Plant* **17**, 725-746 (2024).
24. Liu, Y. *et al.* Pan-genome of wild and cultivated soybeans. *Cell* **182**, 162-176. e13 (2020).
25. Li, Z. *et al.* Gene duplicability of core genes is highly consistent across all angiosperms. *Plant Cell* **28**, 326-344 (2016).
26. Ren, R. *et al.* Widespread whole genome duplications contribute to genome complexity and species diversity in angiosperms. *Molecular Plant* **11**, 414-428 (2018).
27. Feng, X. *et al.* Genomic evidence for rediploidization and adaptive evolution following the whole-genome triplication. *Nature Communications* **15**, 1635 (2024).
28. Wu, S., Han, B. & Jiao, Y. Genetic contribution of paleopolyploidy to adaptive evolution in angiosperms. *Molecular Plant* **13**, 59-71 (2020).
29. Bao, Z. *et al.* Genome architecture and tetrasomic inheritance of autotetraploid potato. *Molecular Plant* **15**, 1211-1226 (2022).
30. Cheng, H., Asri, M., Lucas, J., Koren, S. & Li, H. Scalable telomere-to-telomere assembly for diploid and polyploid genomes with double graph. *Nature Methods* **21**, 967–970 (2024).
31. Healey, A. *et al.* The complex polyploid genome architecture of sugarcane. *Nature* **628**, 804-810 (2024).

ARTICLE IN PRESS

ARTICLE IN PRESS

32. Schreiber, M., Jayakodi, M., Stein, N. & Mascher, M. Plant pangenomes for crop improvement, biodiversity and evolution. *Nature Reviews Genetics* **25**, 563–577 (2024).
33. Hoopes, G. *et al.* Phased, chromosome-scale genome assemblies of tetraploid potato reveal a complex genome, transcriptome, and predicted proteome landscape underpinning genetic diversity. *Molecular Plant* **15**, 520-536 (2022).
34. Small, E. *Alfalfa and relatives: evolution and classification of Medicago*, (NRC Research Press, 2011).
35. Pecrix, Y. *et al.* Whole-genome landscape of *Medicago truncatula* symbiotic genes. *Nature Plants* **4**, 1017-1025 (2018).
36. Kaur, P. *et al.* Delineating the Tnt1 insertion landscape of the model legume *Medicago truncatula* cv. R108 at the Hi-C resolution using a chromosome-length genome assembly. *International Journal of Molecular Sciences* **22**, 4326 (2021).
37. Botkin, J.R., Farmer, A.D., Young, N.D. & Curtin, S.J. Genome assembly of *Medicago truncatula* accession SA27063 provides insight into spring black stem and leaf spot disease resistance. *BMC Genomics* **25**, 204 (2024).
38. Cui, J. *et al.* The genome of *Medicago polymorpha* provides insights into its edibility and nutritional value as a vegetable and forage legume. *Horticulture Research* **8**, 47 (2021).
39. Yin, M. *et al.* Genomic analysis of Medicago ruthenica provides insights into its tolerance to abiotic stress and demographic history. *Molecular Ecology Resources* **21**, 1641-1657 (2021).
40. Wang, T. *et al.* The genome of a wild Medicago species provides insights into the tolerant mechanisms of legume forage to environmental stress. *BMC Biology* **19**, 96 (2021).
41. Li, A. *et al.* A chromosome-scale genome assembly of a diploid alfalfa, the progenitor of autotetraploid alfalfa. *Horticulture Research* **7**, 194 (2020).
42. Chen, H. *et al.* Allele-aware chromosome-level genome assembly and efficient transgene-free genome editing for the autotetraploid cultivated alfalfa. *Nature Communications* **11**, 2494 (2020).
43. Shen, C. *et al.* The chromosome-level genome sequence of the autotetraploid alfalfa and resequencing of core germplasms provide genomic resources for alfalfa research. *Molecular Plant* **13**, 1250-1261 (2020).
44. Long, R. *et al.* Genome assembly of alfalfa cultivar zhongmu-4 and identification of SNPs associated with agronomic traits. *Genomics, Proteomics & Bioinformatics* **20**, 14-28 (2022).

ARTICLE IN PRESS

ARTICLE IN PRESS

45. Cheng, H., Concepcion, G.T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nature Methods* **18**, 170-175 (2021).
46. Rhie, A., Walenz, B.P., Koren, S. & Phillippy, A.M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology* **21**, 245 (2020).
47. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research* **47**, D309-D314 (2019).
48. Christenhusz, M.J., Fay, M.F., Leitch, I.J., of Life, W.S.I.T. & Consortium, D.T.o.L. The genome sequence of spotted medick, *Medicago arabica* (L.) Huds.(Fabaceae). *Wellcome Open Research* **9**, 116 (2024).
49. Ruhsam, M., of Life, W.S.I.T. & Consortium, D.T.o.L. The genome sequence of the Black Medic, *Medicago lupulina* L. *Wellcome Open Research* **9**, 574 (2024).
50. Fornasiero, A. *et al.* Oryza genome evolution through a tetraploid lens. *Nature Genetics* **57**, 1287–1297 (2025).
51. Emms, D.M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology* **20**, 238 (2019).
52. Garrison, E. *et al.* Building pangenome graphs. *Nature Methods* **21**, 2008–2012 (2024).
53. Hickey, G. *et al.* Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biology* **21**, 35 (2020).
54. Zhang, F. *et al.* Evolutionary genomics of climatic adaptation and resilience to climate change in alfalfa. *Molecular Plant* (2024).
55. Waadt, R. *et al.* Plant hormone regulation of abiotic stress responses. *Nature Reviews Molecular Cell Biology* **23**, 680-694 (2022).
56. Sun, H. *et al.* Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. *Nature Genetics* **54**, 342-348 (2022).
57. Kuzmin, E. *et al.* Exploring whole-genome duplicate gene retention with complex genetic interaction analysis. *Science* **368**, eaaz5667 (2020).
58. Wendel, J.F. Genome evolution in polyploids. *Plant Molecular Evolution*, 225-249 (2000).
59. Davydov, E.V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Computational Biology* **6**, e1001025 (2010).

ARTICLE IN PRESS

ARTICLE IN PRESS

60. Walker, B.J., VanLoocke, A., Bernacchi, C.J. & Ort, D.R. The costs of photorespiration to food production now and in the future. *Annual Review of Plant Biology* **67**, 107-129 (2016).
61. Brooks, M.D. & Szeto, R.C. Biological nitrogen fixation maintains carbon/nitrogen balance and photosynthesis at elevated CO2. *Plant, Cell & Environment* **47**, 2178-2191 (2024).
62. López-Calcagno, P.E. *et al.* Overexpressing the H-protein of the glycine cleavage system increases biomass yield in glasshouse and field-grown transgenic tobacco plants. *Plant Biotechnology Journal* **17**, 141-151 (2019).
63. He, F. *et al.* Pan-genomic analysis highlights genes associated with agronomic traits and enhances genomics-assisted breeding in alfalfa. *Nature Genetics* **57**, 1262–1273 (2025).
64. Hämälä, T. *et al.* Impact of whole-genome duplications on structural variant evolution in Cochlearia. *Nature Communications* **15**, 5377 (2024).
65. Vlček, J. *et al.* Whole-genome duplication increases genetic diversity and load in outcrossing *Arabidopsis arenosa*. *Proceedings of the National Academy of Sciences* **122**, e2501739122 (2025).
66. Konečná, V. *et al.* Parallel adaptation in autopolyploid *Arabidopsis arenosa* is dominated by repeated recruitment of shared alleles. *Nature Communications* **12**, 4979 (2021).
67. Baduel, P., Quadrana, L., Hunter, B., Bomblies, K. & Colot, V. Relaxed purifying selection in autopolyploids drives transposable element over-accumulation which provides variants for local adaptation. *Nature Communications* **10**, 5818 (2019).
68. Li, X. & Brummer, E.C. Inbreeding depression for fertility and biomass in advanced generations of inter-and intrasubspecific hybrids of tetraploid alfalfa. *Crop Science* **49**, 13-19 (2009).
69. Dieterich Mabin, M.E., Brunet, J., Riday, H. & Lehmann, L. Self-fertilization, inbreeding, and yield in alfalfa seed production. *Frontiers in Plant Science* **12**, 700708 (2021).
70. Zhang, C. *et al.* Genome design of hybrid potato. *Cell* **184**, 3873-3883. e12 (2021).
71. Cheng, L. *et al.* Leveraging a phased pangenome for haplotype design of hybrid potato. *Nature* **640**, 408–417 (2025).
72. Liu, Z. *et al.* Grapevine pangenome facilitates trait genetics and genomic breeding. *Nature Genetics* **56**, 2804–2814 (2024).
73. Zhou, Y. *et al.* Graph pangenome captures missing heritability and empowers tomato breeding. *Nature* **606**, 527-534 (2022).

ARTICLE IN PRESS

ARTICLE IN PRESS

74. Wang, X. *et al.* Integrative genomics reveals the polygenic basis of seedlessness in grapevine. *Current Biology* **34**, 3763-3777. e5 (2024).
75. Wu, Y. *et al.* Phylogenomic discovery of deleterious mutations facilitates hybrid potato breeding. *Cell* **186**, 2313-2328. e15 (2023).
76. Wang, Y. *et al.* Harnessing clonal gametes in hybrid crops to engineer polyploid genomes. *Nature Genetics* **56**, 1075–1079 (2024).
77. Durand, N.C. *et al.* Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Systems* **3**, 95-98 (2016).
78. Dudchenko, O. *et al. De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92-95 (2017).
79. Alonge, M. *et al.* Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biology* **23**, 258 (2022).
80. Alonge, M. *et al.* RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biology* **20**, 224 (2019).
81. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *Arxiv* **1207.3907** (2012).
82. Marçais, G. *et al.* MUMmer4: A fast and versatile genome alignment system. *PLoS Computational Biology* **14**, e1005944 (2018).
83. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research* **27**, 722-736 (2017).
84. Walker, B.J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS One* **9**, e112963 (2014).
85. Shi, X. *et al.* The complete reference genome for grapevine (*Vitis vinifera* L.) genetics and breeding. *Horticulture Research* **10**, uhad061 (2023).
86. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics* **24**, 637-644 (2008).
87. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236-1240 (2014).
88. Lin, K. *et al.* Beyond genomic variation-comparison and functional annotation of three *Brassica rapa* genomes: a turnip, a rapid cycling and a Chinese cabbage. *BMC Genomics* **15**, 250 (2014).

ARTICLE IN PRESS

ARTICLE IN PRESS

89. Flynn, J.M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences* **117**, 9451-9457 (2020).
90. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0. 2013–2015. (Seattle, USA, 2015).
91. Yan, H., Bombarely, A. & Li, S. DeepTE: a computational method for *de novo* classification of transposons with convolutional neural network. *Bioinformatics* **36**, 4269-4275 (2020).
92. Goel, M., Sun, H., Jiao, W.-B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biology* **20**, 277 (2019).
93. Heller, D. & Vingron, M. SVIM-asm: structural variant detection from haploid and diploid genome assemblies. *Bioinformatics* **36**, 5519-5521 (2020).
94. Jiang, T. *et al.* Long-read-based human genomic structural variation detection with cuteSV. *Genome Biology* **21**, 189 (2020).
95. Smolka, M. *et al.* Detection of mosaic and population-level structural variants with Sniffles2. *Nature Biotechnology* **42**, 1571-1580 (2024).
96. Lovell, J.T. *et al.* GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. *eLife* **11**, e78526 (2022).
97. Wang, Y. *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research* **40**, e49-e49 (2012).
98. Kozlov, A.M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453-4455 (2019).
99. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24**, 1586-1591 (2007).
100. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
101. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100 (2018).
102. Frichot, E., Schoville, S.D., Bouchard, G. & François, O. Testing for associations between loci and environmental gradients using latent factor mixed models. *Molecular Biology and Evolution* **30**, 1687-1699 (2013).

ARTICLE IN PRESS

ARTICLE IN PRESS

103. Fick, S.E. & Hijmans, R.J. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology* **37**, 4302-4315 (2017).
104. Hämälä, T. & Savolainen, O. Genomic patterns of local adaptation under gene flow in *Arabidopsis lyrata. Molecular Biology and Evolution* **36**, 2557-2571 (2019).
105. Hijmans, R.J. *et al.* Package 'raster'. *R package* **734**, 473 (2015).
106. Hijmans, R.J., Phillips, S., Leathwick, J., Elith, J. & Hijmans, M.R.J. Package 'dismo'. *Circles* **9**, 1-68 (2017).
107. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* **14**, R36 (2013).
108. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
109. Putri, G.H., Anders, S., Pyl, P.T., Pimanda, J.E. & Zanini, F. Analysing high-throughput sequencing data in Python with HTSeq 2.0. *Bioinformatics* **38**, 2943-2945 (2022).
110. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550 (2014).
111. Dennis, G. *et al.* DAVID: database for annotation, visualization, and integrated discovery. *Genome Biology* **4**, P3 (2003).
112. Wickham, H. ggplot2. *Wiley Interdisciplinary Reviews: Computational Statistics* **3**, 180-185 (2011).
113. Harris, R.S. *Improved pairwise alignment of genomic DNA*, (The Pennsylvania State University, 2007).
114. Tuskan, G.A. *et al.* The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**, 1596-1604 (2006).
115. Cheng, C.Y. *et al.* Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *Plant Journal* **89**, 789-804 (2017).
116. De Vega, J.J. *et al.* Red clover (*Trifolium pratense* L.) draft genome provides a platform for trait improvement. *Scientific Reports* **5**, 17394 (2015).
117. Schmutz, J. *et al.* Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178-183 (2010).
118. Kreplak, J. *et al.* A reference genome for pea provides insight into legume genome evolution. *Nature Genetics* **51**, 1411-1422 (2019).

ARTICLE IN PRESS

119. Varshney, R.K. *et al.* Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. *Nature Biotechnology* **31**, 240-246 (2013).
120. Bertioli, D.J. *et al.* The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. *Nature Genetics* **48**, 438-446 (2016).
121. Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Research* **14**, 708-715 (2004).
122. Eckes, P., Rosahl, S., Schell, J. & Willmitzer, L. Isolation and characterization of a light-inducible, organ-specific gene from potato and analysis of its expression after tagging and transfer into tobacco and potato shoots. *Molecular and General Genetics* **205**, 14-22 (1986).
123. Lichtenthaler, H.K. & Buschmann, C. Chlorophylls and carotenoids: Measurement and characterization by UV-VIS spectroscopy. *Current Protocols in Food Analytical Chemistry* **1**, F4. 3.1-F4. 3.8 (2001).

Acknowledgements

This work was supported by the Biological Breeding-National Science and Technology Major Project (2022ZD04011) to Q.Y., the National Key R&D Program of China (2025YFF1000700) to R.L., the earmarked fund for CARS (CARS-34) to Q.Y., the Agricultural Science and Technology Innovation Program of Chinese Academy of Agricultural Sciences (ASTIP-IAS14) to Q.Y., the Science Fund Program for Distinguished Young Scholars of the National Natural Science Foundation of China (Overseas) to Y.Z., and the China Postdoctoral Science Foundation (2023M733832) to F.Z.

Authors' contributions

Q.Y., Y.Z., and J.K. conceived and designed the experiments. F.Z., C.W., X.S., S.C., R.L. and J.K. performed the experiments and analysis data. C.W., R.L. and J.K. conducted transgenic experiment. F.Z., X.S., S.C. and R.L. conducted genome assembly and pangenome construction. F.Z., C.W., X.S., Z.W.Z., Y.Z. and Q.Y. wrote the paper. X.X., Z.M., Y.P., R.A., H.X., Z.Z., W.Z., Y.X., Y.D., L.Z., X.C., M.D. and X.W. revised the manuscript. All authors approved the final manuscript.

Competing interests

ARTICLE IN PRESS

The authors declare no competing interests.

ARTICLE IN PRESS

ARTICLE IN PRESS

Table 1. Statistics of the genome assembly and annotation of 13 *Medicago* genomes.

ID	Species	Assembly size (Mb)	Anchored chromosome (Mb)	Contig N50 (Mb)	Number of genes (update/raw) [a]	BUSCO (%) (update/raw) [b]
ZM4_V2.0	subsp. *sativa*	3731.03	3128.69	5.45	202,473/-	99.50/-
Zhongmu1	subsp. *sativa*	816.30	794.30	3.92	52,405/47,780	73.3/71.4
XJDY	subsp. *sativa*	3157.51	2738.00	0.46	170,352/158,894	99.4/97.3
Mca_long	subsp. *caerulea*	801.63	777.90	1.61	50,491/-	96.4/-
Mca_landa	subsp. *caerulea*	793.23	781.07	3.86	53,331/46,579	95.0/91.1
Mt_R108	*M. truncatula*	401.01	391.48	5.93	39,190 /38,449	99.1/97.3
Mt_A17(V5.0)	*M. truncatula*	430.01	426.02	24.88	-/52,135	-/97.2
Mt_HM078	*M. truncatula*	493.76	481.19	31.41	39,527 /37,325	99.7/98.8
Mpoly	*M. polymorpha*	457.53	425.14	11.02	40,070 /35,038	96.5/95.2
Mru_zhiwusuo	*M. ruthenica*	904.13	825.32	0.63	56,497 /45,729	86.5/84.2
Mru_landa	*M. ruthenica*	903.56	828.62	0.68	55,839 /45,530	90.1/89.8
Mara	*M. arabica*	515.95	511.14	6.50	46,150/-	99.01/-
Mlup	*M. lupulina*	575.85	570.12	2.90	47,963/-	98.33/-

[a] The gene annotation was redone for the published genome, excluding contigs. Only the genes present on chromosomes were used for comparison. *M. truncatula* A17 (v5.0) served as the reference database; therefore, it was not utilized for reannotation.

[b] The TEs content percentage was annotated based on the pangenome.

ARTICLE IN PRESS

ARTICLE IN PRESS

Figure legends

Figure 1. Super-pangenome and evolutionary genomics of *Medicago* species. A, Genomic synteny map among the 13 genomes. Different colors represent various chromosomes, with haplotype information for ZM4_V2.0 and XJDY labeled a unique notation ending in 1 to 4. The species names and ploidy information are labeled adjacent to the corresponding genomes. B, Phylogenetic relationships and divergence times among the seven *Medicago* taxa, inferred from 5,106 single-copy orthologous genes. Statistical support for the topology was assessed with 1000 bootstrap replicates. Divergence times were estimated on a fixed phylogeny using a Bayesian MCMC method. The nodes and error bars represent the posterior mean divergence times and the 95% highest posterior density intervals. C, Potential distribution regions of the diploid subsp. *caerulea* (dark blue) and the tetraploid subsp. *sativa* (light blue) predicted by a random forest model. The map is based on bioclimatic data from WorldClim v2[103]. D, The comparison of genome size between the representative genomes and the super-pangenome. E, Correlation between genome size and transposable elements (TEs) content. The *P* value and R^2 value from the linear regression analysis are indicated near the regression lines. The statistical relationship was assessed using a two-sided Pearson's correlation test. Source data are provided as a Source Data file.

Figure 2. The super-pangenome landscape of 13 *Medicago* genomes. A, The bar chart shows the number of gene families in each pangenome category. The pie chart illustrates the pangenome composition. B, Growth curve of pan and core gene families. C, Present and absent information of pan gene families across each genome. D, Pan-genomic features across ZM4_V2.0 hap1 genome. The circular plot displays, from outermost to innermost: the genome (a), density heatmaps for private genes (b), core genes (c), TEs (d), and SVs (e), where darker colors correspond to higher densities. E, Number of structural variations (SVs) across different genomes compared to ZM4_V2.0 hap1. F, Frequency distribution of pangenome SVs across the 13 genomes. G, Principal component analysis (PCA) of the 13 genomes based on pangenome SVs. H, The PCA results of *M. sativa-falcata* complex. Source data are provided as a Source Data file.

Figure 3. The concept of allelic copy number variation and the classification of gene groups. A, A model diagram for identifying allelic variations based on 95% similarity among the same chromosome across four haplotypes. It should be noted that the genes with 1-3 allelic copies are

randomly distributed across 1-3 of the four haplotypes. The percentage numbers indicate the content of each component in the total non-redundant genes. B, Local genomic collinearity among the four haplotypes is depicted, with different colors indicating allelic copy number variation as shown in Figure A. C, Histogram illustrates the allelic copy number variation among different pan-genome categories. D, Selective pressure on genes varies by pan-genome categories and allelic copy number. The sample size (n) for each boxplot is the number of genes within that category. The sample sizes were as follows: Private genes (One copy, n=899; Two copies, n=1,246; Three copies, n=1,141; Four copies, n=860), Dispensable genes (One copy, n=3,193; Two copies, n= 4,118; Three copies, n= 4,066; Four copies, n= 4,120), Softcore genes (One copy, n= 1,479; Two copies, n= 2,243; Three copies, n= 3,790; Four copies, n= 11,844) and Core genes (One copy, n= 1,648; Two copies, n= 2,720; Three copies, n= 5,535; Four copies, n= 24,854). Center line represents the median; box limits indicate the lower and upper quartiles; whiskers represent 1.5 times of the interquartile range; and points beyond this range are outliers. E, Classification of all genes in the genome into six distinct groups, providing both the categorical framework and the genomic baseline proportions for all downstream analyses. Source data are provided as a Source Data file.

Figure 4. The significant contribution of tetra-copy core genes to climate adaptation. A, Histogram illustrates the number of climate-associated genes identified in the LFMM and PBS analysis for each gene group. The red line indicates the enrichment fold. B, Composition of gene groups within each set of differentially expressed genes (DEGs), with each set corresponding to a specific dataset from various stress experiments. The sample size (n) represents the number of independent datasets included in the analysis. The colors for the gene groups are consistent with those in Figure A. C, The gene expression levels of six gene groups. The sample size (n) was as follows: Abiotic stress (Unique-essential, n= 11,233; One-copy, n= 16,029; Two-copy, n= 17,221; Three-copy, n= 20,303; Four-copy, n= 24,190; Tetra-copy core, n= 35,096), Biotic stress (Unique-essential, n= 5,794; One-copy, n= 8,512; Two-copy, n= 9,650; Three-copy, n= 13,419; Four-copy, n= 17,135; Tetra-copy core, n= 27,154), and Normal (Unique-essential, n= 10,458; One-copy, n= 14,994; Two-copy, n= 16,067; Three-copy, n= 19,512; Four-copy, n= 23,368; Tetra-copy core, n= 34,412). Center line represents the median; box limits indicate the lower and upper quartiles; whiskers represent 1.5 times of the interquartile range; and points beyond this range are outliers.

ARTICLE IN PRESS

D, The expression levels of the four allelic copies within each non-redundant gene are arranged in descending order. Data are presented as mean values ± standard deviation (SD), calculated across 16,424 non-redundant tetra-copy core genes. Statistical significance was determined using two-sided pairwise *t*-tests with Bonferroni's correction for multiple comparisons. E, Percentage of different gene groups within top expressed genes across 124 different expression conditions. Data are presented as mean ± SD, calculated across multiple conditions. The sample size (n) was as follows: Abiotic stress (n=63), Biotic stress (n=8), and Normal (n=53). Source data are provided as a Source Data file.

Figure 5. The predominant accumulation of genetic burden in tetra-copy core genes. A, Density distribution of GERP score. Loci with a GERP score >3 were considered deleterious variants (the threshold of the blue dashed line). B, Number of deleterious variants within 1kb regions across different gene groups. Different colors represent various elements of the gene. C, Proportion of non-redundant genes categorized by the number of deleterious allelic copies. D, Gene ontology enrichment results of non-redundant genes carrying four deleterious allelic copies. E, Number of deleterious variants in each accession among subsp. *caerulea* and subsp. *sativa*. The sample size (n) represents the number of accessions. Center line represents the median; box limits indicate the lower and upper quartiles; whiskers represent 1.5 times of the interquartile range; and points beyond this range are outliers. F, Mean number of deleterious variants per gene within each gene group. Both results in E and F were derived from population data using ZM4_V2.0 hap1 as a reference. Deleterious variants were defined as those with a GERP score > 3 and a SIFT4G score < 0.05. Source data are provided as a Source Data file.

Figure 6. The adaptability and genetic burden of the tetra-copy core gene *MsGDC*. A, Local genomic collinearity among seven taxa. Regions highlighted in pink indicate the presence of the *MsGDC* gene. B, Expression patterns of four allelic copies of *MsGDC* under varying nitrogen and CO_2 levels. High_N and Low_N indicate high and low nitrogen levels. Am_CO2 and Ele_CO2 correspond to ambient (420 ppm) and elevated (600 ppm) CO_2 levels. Agate and Saranac are two alfalfa varieties, each of which comprises both effective and ineffective nitrogen-fixing lines (RNA-seq data from Brooks *et al.*[61]). C, Deleterious variants identified in the four allelic copies of *MsGDC*. Yellow, blue, and pink colors represent the 3’ UTR, CDS, and 5’ UTR regions,

ARTICLE IN PRESS

respectively. Each red line indicates a 30 bp region containing one or more deleterious variants. Corresponding gene names are indicated in parentheses. D, Transgenic validation of *MsGDC* in alfalfa. Scale bars are 2 cm (upper) and 1 cm (lower). E, F, Phenotype of *MsGDC*-overexpressing transgenic *M. truncatula* before (E) and after (F) nitrogen-free treatment. Scale bars, 2 cm. For panels D-F, WT represents wild type. G lines (MsG in alfalfa, MtG in *M. truncatula*) are transgenic lines using the 35S promoter. GS lines (MsGS in alfalfa, MtGS in *M. truncatula*) use the ST-LS1 promoter. Source data are provided as a Source Data file.

Figure 7. Model of evolutionary consequences of autopolyploidy: impacts on adaptive advantages and long-term evolutionary fate. Schematic of a pangenome comparing four distinct species (A-D). Gene categories (core, softcore, etc.) are color-coded, and deleterious variants are marked by red bars. Following WGD, a subset of core genes were retained in four allelic copies (tetra-copy core genes, highlighted by a yellow background and enclosed in a red dashed box), and these genes are shown to be significant contributors to climate adaptation. For each gene set, one allelic copy is primarily responsible for climate adaptation roles. The other three allelic copies are hypothesized to modulate the expression of the functional copy. However, these adaptive advantages may come at a long-term cost. The process of adaptation can lead to the accumulation of deleterious variants, a phenomenon known as genetic burden. Over time, an excessive accumulation of this burden within the tetra-copy core genes could ultimately compromise the fitness of autopolyploids, potentially leading them to an evolutionary 'dead end'. Schematic elements representing environmental stress were created in BioRender. Wang, X. (2025) https://BioRender.com/66lnuem.

ARTICLE IN PRESS

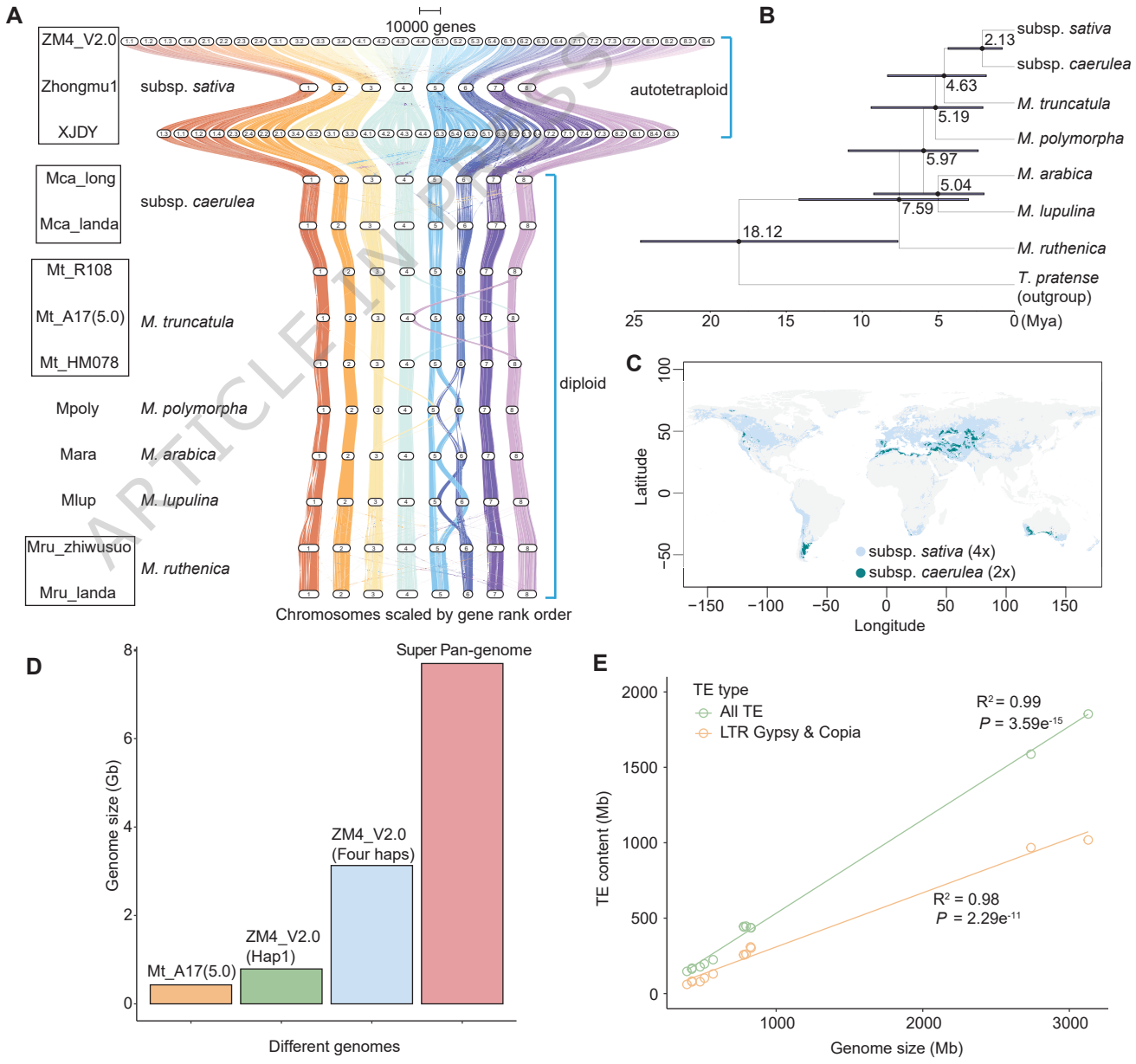
ARTICLE IN PRESS

ARTICLE IN PRESS

Editor's Summary

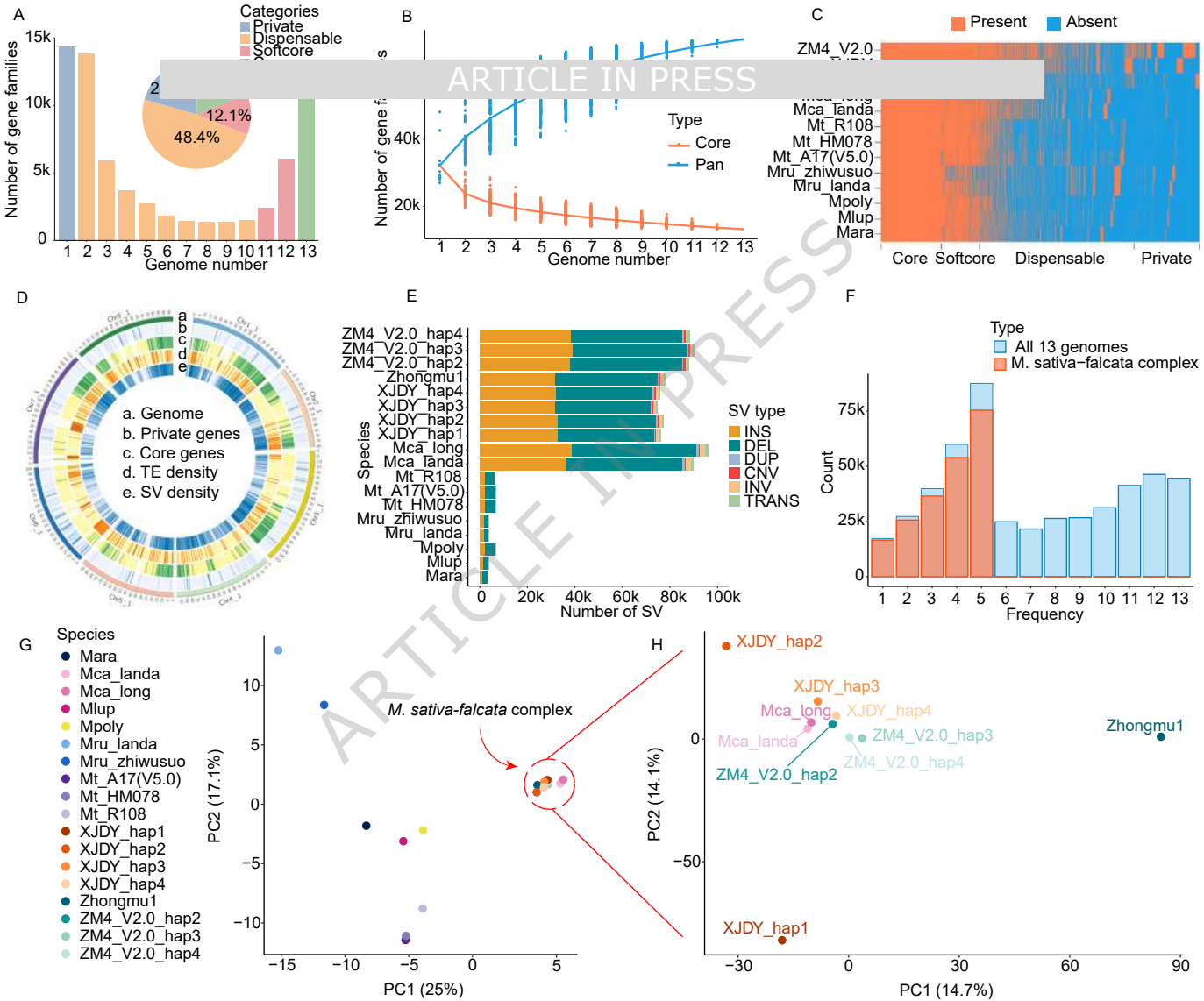
The genetic mechanisms of why polyploid plants exhibit adaptive advantages over diploids are unclear. Here, the authors assemble the autotetraploid alfalfa genome, incorporate it into a pangenome analysis, and reveal the trade-off between adaptive advantages and evolutionary constraints mediated by tetra-copy core genes.

Peer review information: *Nature Communications* thanks Yong Zhou and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

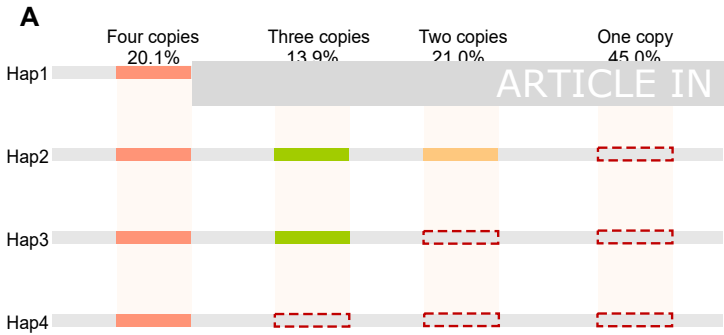


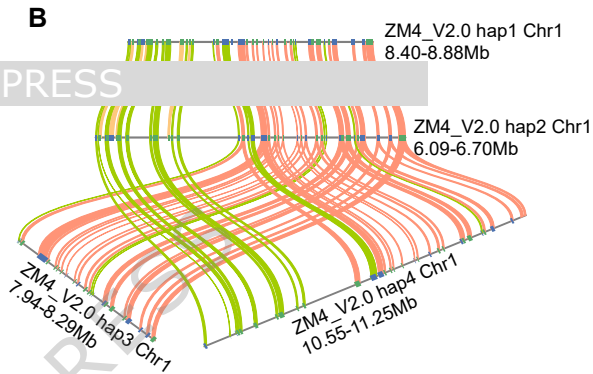
ARTICLE IN PRESS

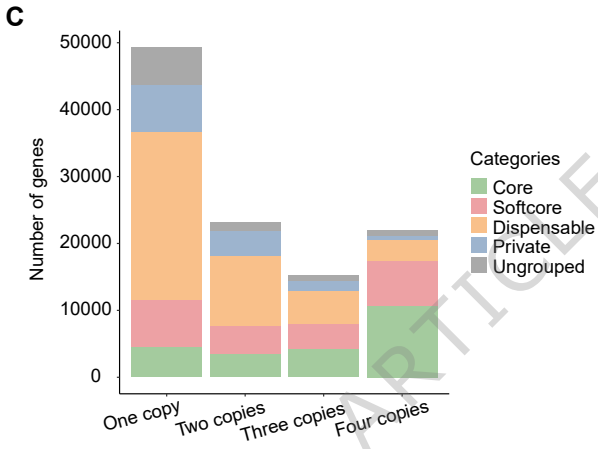
ARTICLE IN PRESS

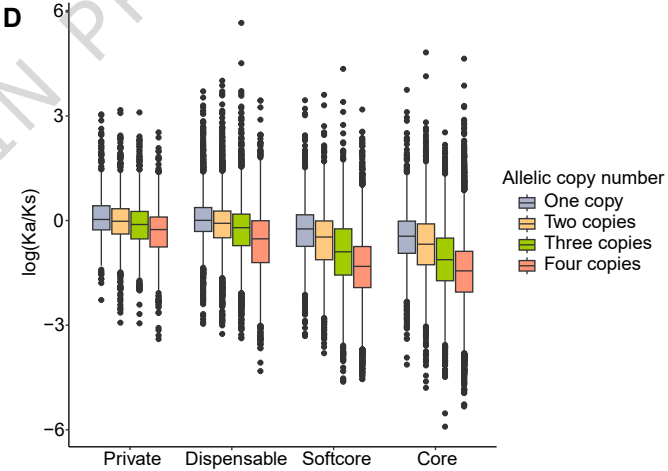


ARTICLE IN PRESS





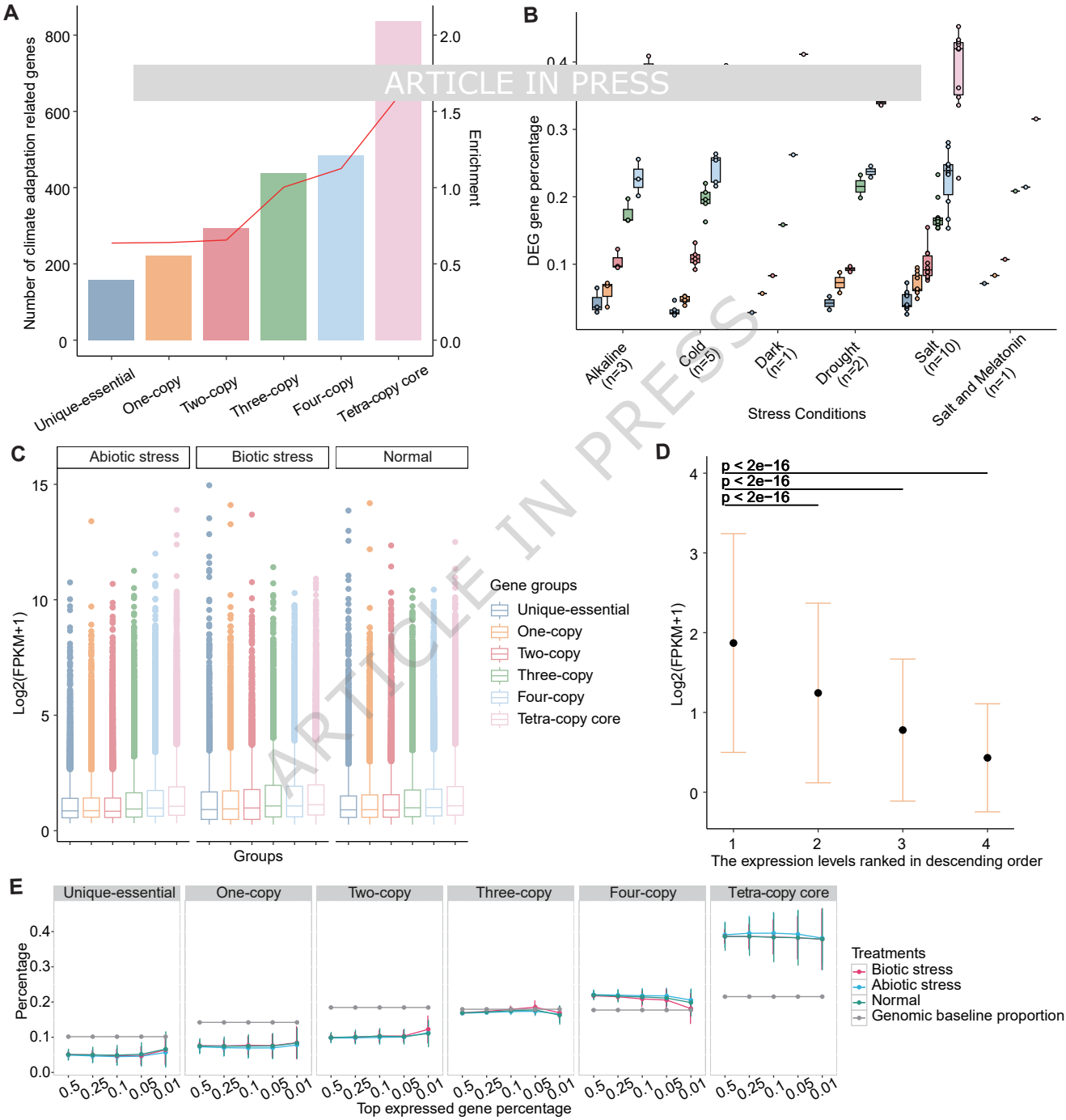




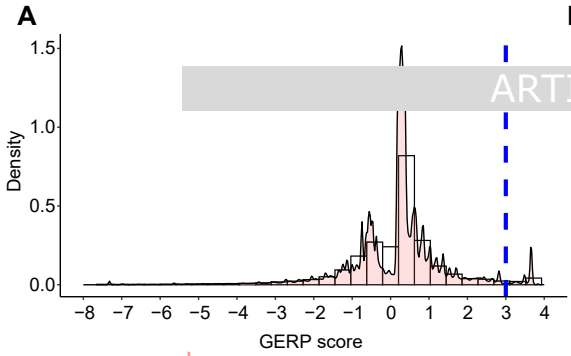
E

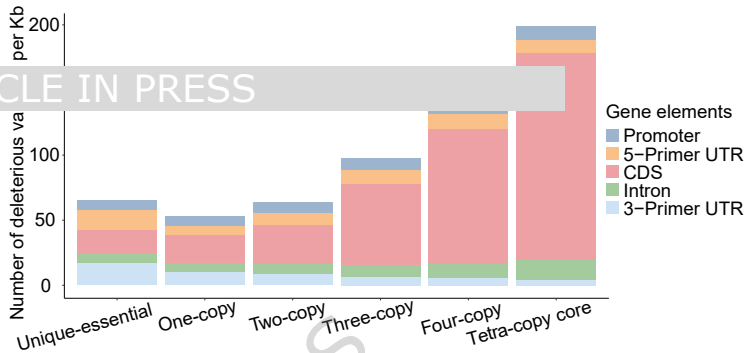
Allelic variations \ Categories	Core	Softcore	Private	Ungrouped	Dispensable
Four copies	Tetra-copy core genes (21.5%)	Four-copy genes (17.7%)			
Three copies	Three-copy genes (17.9%)				
Two copies	Two-copy genes (18.4%)				
One copy	One-copy genes (14.2%)				Unique essential genes (10.2%)

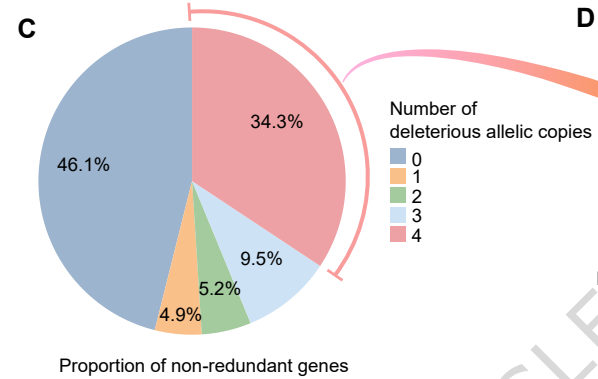
ARTICLE IN PRESS

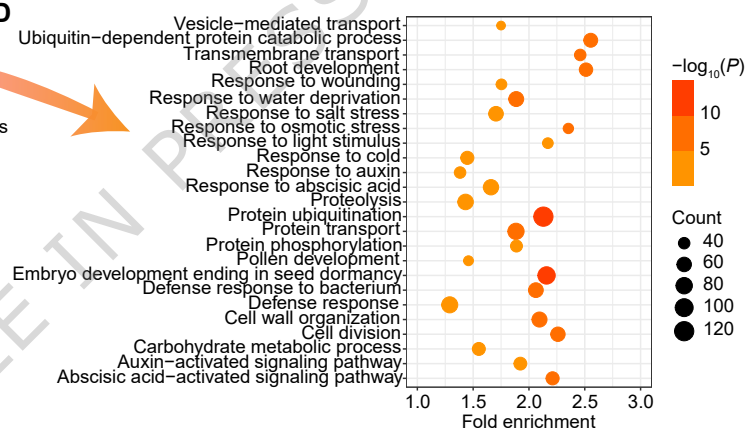


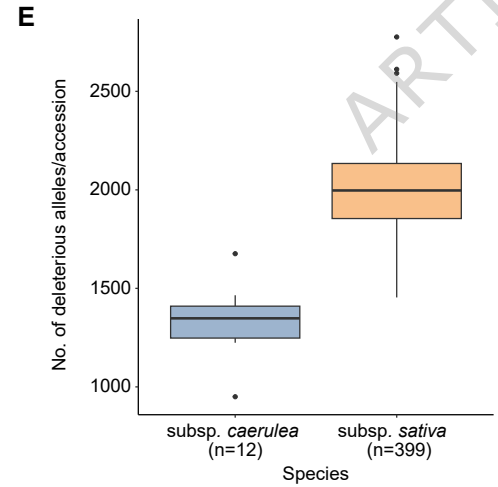
ARTICLE IN PRESS

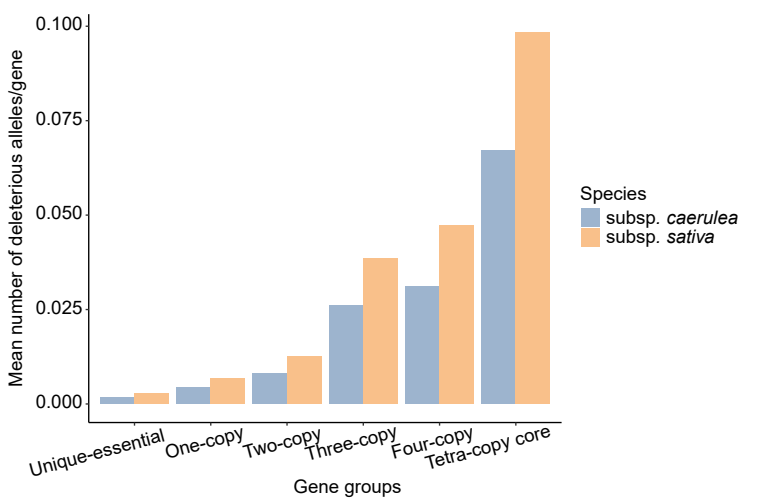




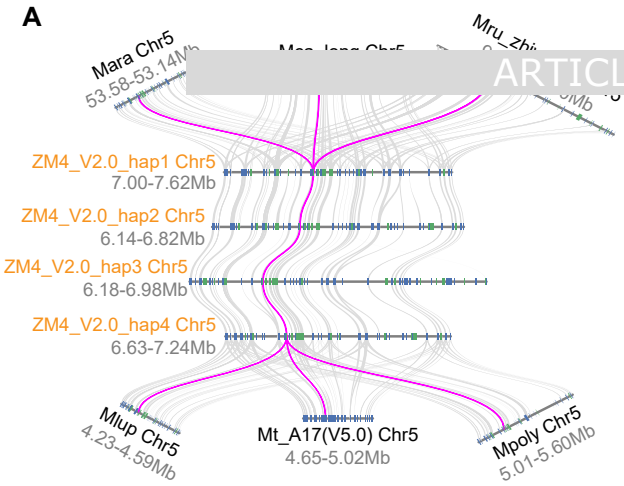


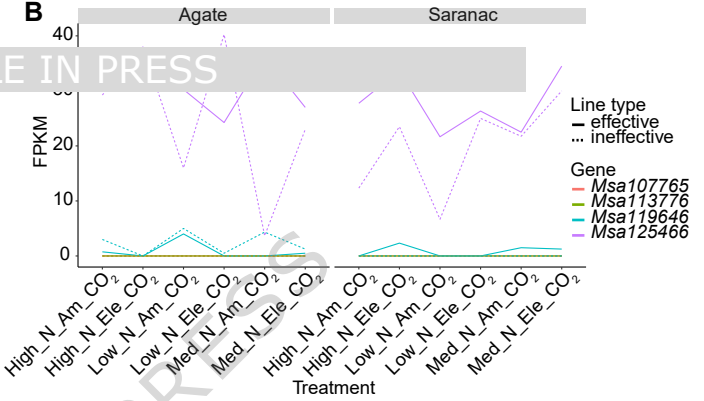


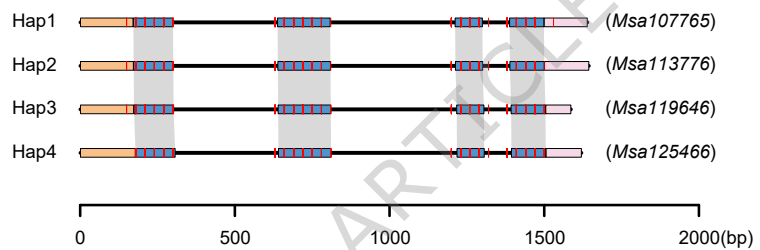


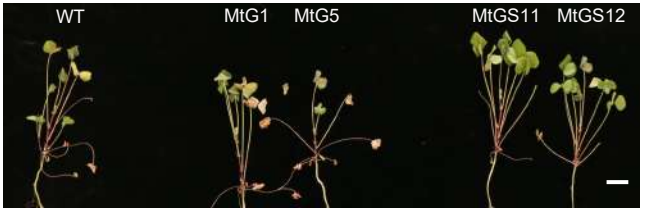


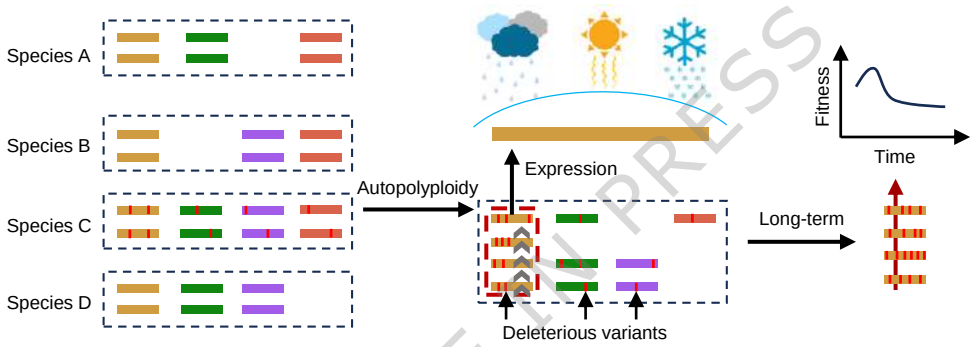
ARTICLE IN PRESS











ARTICLE IN PRESS